# Characterizing Debate Performance via Aggregated Twitter Sentiment

**Nicholas A. Diakopoulos**
Rutgers University
School of Communication and Information
diakop@rutgers.edu

**David A. Shamma**
Yahoo! Research
Internet Experiences
aymans@acm.org

**ABSTRACT**
Television broadcasters are beginning to combine social micro-blogging systems such as Twitter with television to create social video experiences around events. We looked at one such event, the first U.S. presidential debate in 2008, in conjunction with aggregated ratings of message sentiment from Twitter. We begin to develop an analytical methodology and visual representations that could help a journalist or public affairs person better understand the temporal dynamics of sentiment in reaction to the debate video. We demonstrate visuals and metrics that can be used to detect sentiment pulse, anomalies in that pulse, and indications of controversial topics that can be used to inform the design of visual analytic systems for social media events.

**Author Keywords**
Video, TV, Affect, Twitter, Annotation, Debate, Journalism

**ACM Classification Keywords**
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

**General Terms**
Human Factors, Design

## INTRODUCTION

In the fall of 2008, Current TV ran a program called *Hack the Debate* where they called for people to microblog comments during a live event. Using the popular Twitter service, these posts—called *tweets*—were displayed on TV underneath the live presidential debate between Barack Obama and John McCain. The success of Current's program has lead to many broadcasters to call for tweets during live broadcasts. While viewers can see opinions one by one when watching, the collection of tweets provides an opportunity to understand the overall sentiment of microbloggers during the event.

Twitter is a microblogging platform that limits each post to 140 characters, which is slightly less than an SMS/text message to a cell phone. Similarly, it is just text and does not support other formats like pictures or videos; people add URLs to their posts when they wish to send rich media. Each user updates their feed. A user can also watch, or *follow*, another user's feed. This creates a publish-and-subscribe social network where each user has a followers and following count. Additionally, within a tweet, a specific twitter user can be mentioned by prefacing their username with an @ symbol. This creates links between users and allows for threaded conversations between users.


*CHI 2010*, April 10−15, 2010, Atlanta, Georgia, USA.


When people tweet live about a media event they are in effect annotating. When mined for their affective content, these annotations can identify parts of the video that gained interest or proved controversial. In this work we wish to characterize a media event, a debate, according to how people are reacting to it. We are however not interested in the automatic detection of a winner or loser. To do this, we describe an analytic methodology for detecting affective patterns that could aid in the development of media analytical tools. Such a tool could serve to help a journalist or public affairs person become aware of trends and patterns in public opinion around media events.

## RELATED WORK

Several prior studies have examined the usage patterns and the social motives of Twitter, such as through metrics of reciprocity. Java et al. [5] compared micro-blogging versus regular blogging. They found users of micro-blogging systems to engage in a higher social reciprocity as measured by a publish-to-subscribe ratio. Krishnamurthy et al. [7] measured the same reciprocity but added for posting frequencies. Honeycutt and Herring [4] measured the usage of the @ symbol to measure conversational engagement. Naaman et al. [9] coded a sample of tweets to broadly classify users as self-broadcasters or informers. Our work builds on that of Shamma et al. [12] who began to study tweets relating to media events. By examining conversation volume and activity over time, they were able to temporally segment a live news event and identify the key people in the event.

Explicit media annotation and sharing while watching TV has been studied in a variety of manners [3, 15] however these systems often involve integrated set-top boxes to support collaboration. Online video annotation and conversation has also been examined, such as the Videolyzer system which supports collaborative information quality annotation of video [1]. Nakamura et. al. [10] have studied affective response on unstructured video commenting systems such as the popular Japanese video site *NicoNicoDouga*. Other work has more broadly characterized temporal patterns of messaging behavior on social networks, though not in conjunction with sentiment or with anchor media [2]. This

work focuses on using the sentiment of tweet annotations to understand their relationship to topicality, as well as to the rhythm and performance of actors, in this case presidential candidates, in the event.

## EVALUATIVE TWEET STUDY

To study the tweets about the debate, we crawled the Twitter search API for common related tweets by looking for related *hashtags*. The mechanism for tagging posts on Twitter relies on the poster to prefix a term with the # symbol. For the first presidential debate of 2008, we queried the Twitter Search API for #current, #debate08 and #tweetdebate. This amounted to 1,820 tweets from 664 people during the 97-minute debate and 1,418 tweets from 762 people in the 53 minutes following the debate. During the debate there was an average of 2.74 messages per user (Median = 1, SD. = 4.54). The top contributor made 42 tweets and 5.7% of users made 10 or more tweets indicating that there was a diverse distribution of user activity with most people chiming in only a single time.

### Measuring Tweet Sentiment

#### *Rating Acquisition*

In order to understand the valence of the sentiment during the debate we collected three independent sentiment ratings for each of the 3,238 tweets in our corpus. Tweets were rated as belonging to one of four categories: *negative*, *positive*, *mixed*, and *other*. "Mixed" tweets included those that contained both positive and negative components and "other" was a category included to catch non-evaluative statements or questions.

Ratings were acquired using Amazon Mechanical Turk (AMT), a crowd-sourcing site where workers complete short tasks for small amounts of money. AMT ratings have been shown in prior linguistic rating experiments to correlate well with and sometimes outperform expert raters [14]. Workers were compensated $0.05 for each batch of ten ratings that they submitted.

As there is oftentimes noise in AMT ratings [6], we applied a total of five filters to enhance the overall quality of the ratings and discard ratings from workers suspected of poor quality ratings. The first filter was a time filter in which a batch of ten ratings was discarded if the amount of time it took the worker to submit those ratings was less than one standard deviation below the mean submission time for all workers. Next we applied a sloppiness filter: if a worker did not submit a rating for any of the ten tweets in a batch, then all ten ratings were discarded; we might infer the worker was not being careful or thoughtful. Each of the batches contained one from a set of control tweets that had an obvious and verified sentiment. If a worker mislabeled a control tweet we discard all ten ratings in that batch.

We also included a simple worker bias and an overall worker quality filter. The worker bias filter operates by measuring the distribution of ratings across the four categories for each rater. If the ratio of positive to negative ratings was above a threshold or below another threshold for a user, we infer the user is biased in one direction or another. More sophisticated bias correction schemes have also been developed [14], but we found this simple filter eliminated ratings from the most blatantly biased workers.

The overall worker quality filter works by discarding the remaining ratings from workers whose ratio of ratings retained to ratings already discarded is below 0.5. This threshold guarantees that in aggregate the quality of the ratings will improve [13]. Intuitively, if someone has more than half of their ratings discarded from the filters they are likely a poor rater and we discard their remaining ratings. Using these five filters we discarded 60% of all of the ratings collected from AMT.

#### *Rating Results*

Since our rating categories are not mutually exclusive a rating reliability measure such as Cohen's Kappa or Fleiss' Kappa is not appropriate. We adopt a technique from [14] which computes the inter-annotator agreement (ITA) as the average Pearson correlation for each set of ratings with the aggregate rating. The aggregate rating was produced for each tweet using a simple majority-voting rule over the three independent ratings. Correlations were averaged across all possible ways to break ties in cases where there was no consensus. Using this method we achieved an ITA of .655, indicating a good amount of agreement between ratings. In total, 1,187 tweets (36.7%) had perfect agreement among the three raters, 1,622 tweets (50.1%) had consensus from two of the three raters, and 429 tweets (13.2%) had no consensus in labeling.

To verify that these ratings were accurate we had three experts (the two authors plus one other colleague) rate a subset of 200 randomly chosen tweets from the dataset. The ITA for these expert ratings was 0.744, indicating that experts still agree with experts more often than non-experts agree with non-experts. However, we believe our aggregated non-expert ratings are still adequate for drawing some conclusions about the sentiment response to the debate.

### Characterizing Tweets During the Debate

In this section we utilize the aggregated tweet ratings to characterize the debate in terms of the overall sentiment of the tweets, whether twitter users favored a particular candidate, and the temporal evolution and "pulse" of the sentiment observable in the tweets. We were also interested in being able to detect anomalies in this pulse as well as understand the relationship of sentiment to the topicality and potential controversy of issues being discussed. The overarching goal of this characterization was to understand what features would lend themselves toward a temporal media event analysis system as might be employed by a journalist or public affairs person.

#### *Sentiment and Favor*

The tenor of the tweets during the debate was distinctly negative (41.7% of tweets). Positively tagged tweets represented 25.1% of the set, mixed tags accounted for 6.8% and the remainder of 26.4% consisted of tweets tagged as

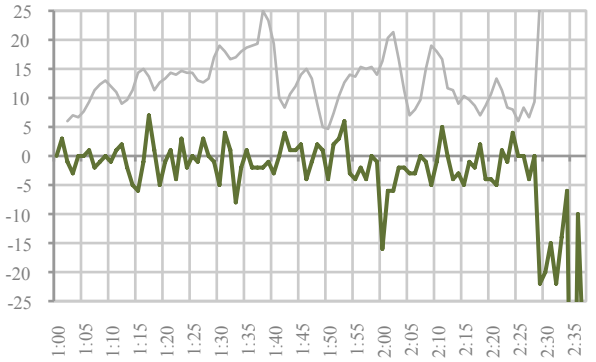


**Figure 1. The number of positive minus the number of negative tweets per minute, with 3 minute moving average of total tweet volume in light gray. Times in GMT.**

"other" or for which there was no tagging consensus. The overall dominant negative response is consistent with theories of negativity in political evaluation formulation [8].

To understand whether tweet sentiment was favoring one candidate or another we used C-SPAN's (a news agency) transcript and timing information as metadata for who was speaking during each minute of the debate—Barack Obama, John McCain, or the moderator Jim Lehrer. For each minute we also define the aggregate valence of response as the number of positive tweets minus the number of negative tweets. We excluded minutes from our analysis where both candidates spoke substantially since that would conflate response scores. For minutes when only Obama spoke, the mean aggregate valence score was -2.09; for minutes when only McCain spoke the mean aggregate valence score was -5.64. The sentiment of tweets suggests that tweeters favored Obama over McCain, with McCain's aggregate valence more than twice as negative as Obama's.

### *Sentiment Evolution and Pulse*

The evolution of the valence of the tweets over the course of the 97-minute debate can be seen in Figure 1. The aggregate valence of the debate fluctuated with who was speaker at that particular time and the overall valence declined and then fell steeply during the last 10 minutes. Examination of individual tweets during this final period indicates that a combination of both the impending end of the event together with an inciting topic (terrorism) led to a higher volume of activity.

To understand the pulse and periodicity of the aggregate valence shifts we took the discrete Fourier transform and found that the dominant frequency in the signal corresponds to a period of 5.19 minutes. This is the amount of time it took for both candidates to take a complete turn and can for example be seen quite pronouncedly between minutes 12 and 18 in Figure 1. Looking at the individual tweets during this period we confirmed that the peak valence response corresponds to when Obama was speaking and the trough response to when McCain was speaking.

The other peaks and valleys in Figure 1 can be used to identify areas of the debate where either candidate was getting

**Table 1. A timeline of the debate showing Pearson correlation scores of positive and negative tweets by topic.**

| GMT | Topic | Correlation | P-Value |
|---|---|---|---|
| 1:01:34 | Opening | 0.051 | > 0.2 |
| 1:03:12 | Financial Recovery | 0.623 | < 0.05* |
| 1:14:06 | Solving Financial Crisis | -0.470 | < 0.15 |
| 1:26:00 | Financial Recovery | 0.528 | < 0.05* |
| 1:38:54 | Lessons of Iraq | 0.229 | > 0.2 |
| 1:50:11 | Troops in Afghanistan | -0.142 | > 0.2 |
| 2:03:11 | Threat from Iran | 0.313 | > 0.2 |
| 2:15:47 | Relations with Russia | -0.188 | > 0.2 |
| 2:25:53 | Terrorist Threat | 0.662 | < 0.02* |

their "expected" response—either positive or negative. However an analyst interested in the performance of the debaters might also be interested in when the pulse was disrupted—anomalies when either candidate was underperforming or over performing as compared to their average aggregated valence response. In some cases this would correspond to "flat" areas of Figure 1.

To help detect these anomalous areas we plot how much the aggregate valence score differs from the mean aggregate valence score for that candidate in Figure 2. Looking at Figure 2 we can for instance quickly see that minute 17 was an exceptionally strong moment for Obama and that McCain had a strong point at minute 53. We can also see weakness for Obama at minutes 56–57, and comparative strength for McCain at minutes 58–59 followed by an exceptionally weak point for McCain at minute 60.

The period between minutes 53–60 is a bit different in its signature so we looked to individual tweets to help explain the pattern at that time. The candidates were addressing military issues—in particular troops in Afghanistan—at that time. This seemed to bring out more positive reactions for McCain such as "*You have to admit, McCain is VERY knowledgeable about foreign policy & what's happening in the middle east.*" At minute 60 McCain tells an emotional war story for which the tweets are resoundingly critical.

### *Controversial Segments*

The moderator broke this debate into distinct topics, which we collected from C-SPAN's website (see Table 1). In order to give some indication of controversy we computed the Pearson correlation between the positive and negative responses for each topic. Intuitively, a high correlation indicates that the given topic arouses interest on both sides of the issue according to some consistent pattern. However, without a deeper examination of the tweets themselves, the correlation only *suggests* controversy since factors such as interest level (i.e. total volume of tweets) as well as the balance between positive and negative response (i.e. interest from both sides) are also important.

We did find significant correlations between positive and negative sentiment on the topics of *financial recovery* and *terrorist threat*. For the first section on financial recovery this signature is also observable in Figure 1 as a relatively flatter curve between minutes 3–14. The financial recovery

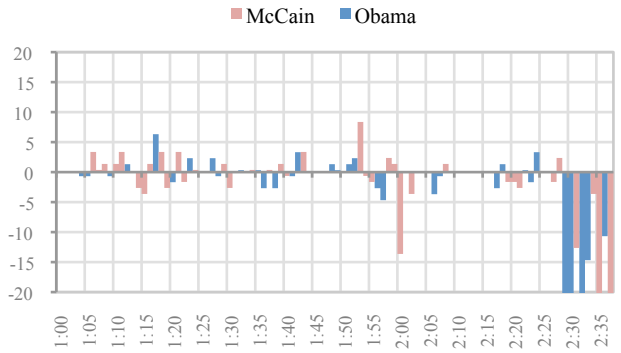


**Figure 2. Strength of response compared to mean valence for minutes when that candidate spoke. Times in GMT.**

section was also broken into two pieces, separated by a segment on solving the financial crisis. The positive and negative sentiment was correlated in both segments, despite the uncorrelated 12-minute topic shift in-between. Looking at the volume curve in Figure 1 we can also see that the second section on financial recovery as well as the section on terrorist threat are areas of high message volume.

## DISCUSSION

One of the issues with this form of event annotation is that it *infers* a relationship between a media event and an affective response via a timestamp and a hashtag. In reality, there were tweets during the debate which were evaluative, but which did not reference the event itself. For instance, someone might be critical in response to another commenter or about something that is irrelevant to that *particular* time of the event. In the future we intend to add more detailed textual analytics to help the analyst further disambiguate the Twitter response.

The debate tweets do not represent everyone who watched the debate, only those who had adopted Twitter and had chosen to respond. Measuring population sentiment from a system like Twitter could not be substituted for a real poll. As real-time social commenting around media events becomes more prevalent and the biases of users of these systems tend toward population biases, it will be helpful to have knowledge about the background of a user, such as political leaning or even just age, in order to better see the sentiment response of different slices of users. While some of this will be explicitly available from user profiles, future work could also look at inferring background from sentiment response. For example, can we predict a user's political leaning based on the history of their sentiments during either candidate's speaking minutes?

Since the response time for something like Mechanical Turk would be too long for a journalist trying to make sense of an event in near real-time, the analytic methodology that we have developed will require automatic methods for classifying tweets into positive and negative sentiment. Using our annotated data as a training set we are confident that known automatic techniques for sentiment classification [11] can achieve viable results for such an application.

## CONCLUSIONS

We have demonstrated an analytical methodology including visual representations and metrics that aid in making sense of the sentiment of social media messages around a televised political debate. We demonstrated that the overall sentiment of the debate was negative and that tweeters tended to favor Obama over McCain. We also showed that interesting events can be detected by looking at anomalies in the pulse of the sentiment signal and that controversial topics can be identified by looking at correlated sentiment responses. This analysis is highly dependent on the polarized structure of a political debate, however we wish to explore how other events, (speeches, TV shows, sports), could also be analyzed using sentiment. We suggest that a system embedding such metrics and visuals as we have developed here could enable journalists to identify key sections of a debate performance, or could enable public affairs officials to optimize a candidate's performance.

## REFERENCES


1. Diakopoulos, N., Goldenberg, S. and Essa, I., Videolyzer: Quality Analysis of Online Informational Video for Bloggers and Journalists. in *Proc. CHI*, (2009).
2. Golder, S., Wilkinson, D. and Huberman, B., Rhythms of social interaction: messaging within a massive online network. in *Communities and Technologies*, (2007).
3. Harboe, G., Metcalf, C.J., Bentley, F., Tullio, J., Massey, N. and Romano, G., Ambient social TV: drawing people into a shared experience. in *CHI*, (2008), 1-10.
4. Hunneycutt and Herring, Beyond Microblogging: Conversation and Collaboration via Twitter. in *HICSS*, (2009).
5. Java, A., Song, X., FInin, T. and Tseng, B., Why we twitter: understanding microblogging usage and communities. in *Workshop on Web Mining and Social Network Analysis*, (2007).
6. Kittur, A., Chi, E.H. and Suh, B., Crowdsourcing User Studies With Mechanical Turk in *Proceedings of CHI*, (2008), 453-456.
7. Krishnamurthy, B., Gill, P. and Arlitt, M. A few chirps about twitter *Workshop on online social networks (WOSP)*, 2008.
8. Lau, R.R. Negativity in Political Perception. *Political Behavior*, *4* (2). 353-377.
9. Naaman, M., Boase, J. and Lai, C.-H., Is it Really About Me? Message Content in Social Awareness Streams. in *CSCW*, (2010).
10. Nakamura, S., Shimizu, M. and Tanaka, K., Can Social Annotation Support Users in Evaluating the Trustworthiness of Video Clips? in *Workshop on Information Credibility on the Web (WICOW)*, (2008), 59-62.
11. Pang, B. and Lee, L. *Opinion Mining and Sentiment Analysis*, 2008.
12. Shamma, D.A., Kennedy, L. and Churchill, E. Tweet the debates *ACM Multimedia Workshop on Social Media (WSM)*, (2009).
13. Sheng, V.S., Provost, F. and Ipeirotis, P.G., Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. in *KDD*, (2008).
14. Snow, R., O'Connor, B., Jurafsky, D. and Ng, A.Y., Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. in *EMNLP*, (2008).
15. Williams, D., Ursu, M.F., Cesar, P., Bergström, K., Kegel, I. and Meenowa, J., An emergent role for TV in social communication. in *European Interactive Television Conference (Euro ITV)*, (2009), 19-28.