Diamonds in the Rough: Social Media Visual Analytics for Journalistic Inquiry

Nicholas Diakopoulos, Mor Naaman, Funda Kivran-Swaine

Rutgers University, School of Communication and Information

ABSTRACT

Journalists increasingly turn to social media sources such as Facebook or Twitter to support their coverage of various news events. For large-scale events such as televised debates and speeches, the amount of content on social media can easily become overwhelming, yet still contain information that may aid and augment reporting via individual content items as well as via aggregate information from the crowd's response. In this work we present a visual analytic tool, Vox Civitas, designed to help journalists and media professionals extract news value from large-scale aggregations of social media content around broadcast events. We discuss the design of the tool, present the text analysis techniques used to enable the presentation, and provide details on the visual and interaction design. We provide an exploratory evaluation based on a user study in which journalists interacted with the system to explore and report on a dataset of over one hundred thousand twitter messages collected during the U.S. State of the Union presidential address in 2010.

KEYWORDS: Computational Journalism, Computer Assisted Reporting, Social Media, Sensemaking

INDEX TERMS: H.5.2 Information Interfaces and Presentation: User Interfaces

1 INTRODUCTION

Social media systems have proven to be valuable platforms for information and communication, in particular during events such as the magnitude 7.0 earthquake that rattled Haiti in 2010. In recognition of this phenomena, journalists are increasingly turning to social media sources like Twitter, Facebook, or other online sources of user content in an effort to track the importance of stories and to find sources of expertise to drive new stories [17]. However, the rush of information from millions of new "human sensors" contributing information about events and news stories leads to the challenge of making sense of the overwhelming response, both at an individual and aggregate level of analysis.

In this paper, we consider how social media content contributed around large-scale broadcast news events can inform journalistic inquiry. For instance: what kinds of insights, analyses, and other activities can be enabled through the support of visual analytic tools in the context of journalism? In particular, we designed and evaluated a visual analytics system, Vox Civitas, whose goal is to make the social media (e.g., Twitter) response to events more amenable to journalistic investigation and sensemaking.

diakop@rutgers.edu
mor@rutgers.edu
funda@eden.rutgers.com

IEEE Symposium on Visual Analytics Science and Technology
October 24 - 28, Salt Lake City, Utah, USA
978-1-4244-9486-6/10/$26.00 ©2010 IEEE

Journalistic investigation and sensemaking poses a somewhat different context than the more oft studied context of investigative analysis for intelligence [2, 13, 19, 20]. Such studies in the intelligence analysis context are certainly valuable for informing the general cognitive processes involved with information analysis. Here, we are more concerned with the context of journalism, including the goals and work products journalists are tasked with. Through the design and evaluation of Vox Civitas we are exploring the domain of *journalistic* analysis in response to social media data, including implications for the design of appropriately tailored visual analytics tools.

We were careful in our design of Vox Civitas to consider journalistic and news values [14] and use these to inform the filtering capabilities and visual schema that were chosen to organize the information stream. In addition to the value-sensitive design rationale that we provide for Vox Civitas, we contribute results of an exploratory study that assessed the utility of the tool. The study, using a dataset from the Twitter response to the U.S. State of the Union presidential address in 2010, examined the kinds of journalistic activities the application supports, and the ways in which the schema and features designed into the application are used by journalists in this context. We relate our findings to the sensemaking model of Pirolli and Card [20]. Understanding if, how, and why the features designed into Vox Civitas support journalistic inquiry will inform the design of future systems built for journalistic sensemaking activities.

2 RELATED WORK

Our work is most inspired by the work of Shamma et al. [24, 25] who have looked at revealing (and to a more limited extent visualizing) the structure and dynamics of twitter content around broadcast media events such as the Presidential Debates. In their work, the authors identify usage cues (magnitude of response) and content cues (salient keyword extraction) as indicators of interesting occurrences in the event such as topic shifts. Our work builds on these ideas in several important ways by integrating such usage and content cues with powerful filtering and interaction mechanisms, derivative data facets such as sentiment, and visual methods for schematizing analyses for journalistic purposes. Moreover, we present an exploratory evaluation of our system in the context of journalistic sensemaking.

Other related work has examined social media content as an information source in the context of emergency response and crisis scenarios such as fires [5] and floods [26]. Indeed, Starbird et al.'s [26] study of the Twitter response to the Red River flooding in early 2009 showed that Twitter users are participating in useful information generation and synthesis activities but are part of a larger ecosystem involving information from traditional media outlets. It is the generative and synthetic activity of social media users that we hope to harness in the context of visual analytics for journalism. Our long-term goal is to enable traditional media to go beyond simply publishing in social media platforms and harness it to drive new insights and stories leading to a virtuous cycle of collaborative sensemaking between social media participants and the newsroom.

The analysis of text corpora over time has been addressed by a variety of systems including ThemeRiver [12], which looks at the evolution of topics over time; Narratives [8], which allows users to track, analyze, and correlate the blog response to news stories over time; and MemeTracker [15] which visualizes the patterns of phrases that appear in news and social media content over time. Recent research has also looked at assessing thematic story visualization in the context of dynamically evolving information *streams* [22]. Our approach differs insofar as thematic change in social media is not the analytic end goal but rather an input in a matrix of analytic enablers including sentiment analysis and journalistically motivated data filters.

The visual analysis of sentiment in large text corpora has garnered some attention in the visualization literature. Gregory et al. [10] presented the integration of sentiment analysis visualizations into the IN-SPIRE system though there was no attempt at temporal analysis. Wanner et al. [27] looked at visualizing sentiment trends in streams of RSS news feeds around the U.S. presidential election in 2008. Diakopoulos and Shamma [7] presented temporal visuals which depict sentiment patterns (e.g. periodicity, strength or weakness of actors) in the context of the social media response to the U.S. presidential debates. Here, we go beyond these prior systems to visually connect sentiment patterns with topicality and the magnitude of the response.

3 Designing For Journalistic Inquiry

In this section our intent is to describe some of the aspects of journalistic practice and values that inform the design of Vox Civitas. In particular, our design objective is to be able to direct attention to the pieces of information that may be most interesting journalistically, as well as to schematize the visual representations in ways that enable improved journalistic inquiry. And while many of the design ideas stem from normative descriptions of journalistic practice, the development of Vox Civitas also involved iterative gathering of feedback from several journalists.

Journalism can be defined as the professionalized practice of "producing and disseminating information about contemporary affairs of general public interest and importance" [23]. Journalistic values include notions of accuracy, balance, and objectivity as well as a keen emphasis on telling an engaging and clear story using primary source interview quotations [14].

What are the types of questions that journalists would reasonably ask of a social media visual analytics tool? We can look to how media events have been covered by the news in the past for indicators of importance and newsworthiness that would inform the design of our system. For instance, studies of newspaper coverage of televised political messages such as debates have shown a tendency for the news to cover "decisive moments" with a preference for moments of clearly divergent points of view, where the audience has indicated approbation or criticism, and which are easily extractable and can stand on their own with minimal need for re-contextualization [3]. Moreover, general newsworthiness guidelines in journalism tend to favor stories that are in some way surprising, unusual, or which are particularly good or bad news [11]. These findings and values imply that sentiment analysis (applause, criticism, controversy, good, bad) in conjunction with the magnitude of the social media response to different quotes, topics, or issues in the speech will be useful analytic indicators for journalistic inquiry.

In order to initially assess what the signal to noise ratio of useful information sharing is on social network systems like Twitter, we collected a sample of ~900 twitter messages (tweets) made in response to Obama's speech at the Copenhagen 15 meeting in December 2009. Qualitative analysis of this (albeit small) sample confirmed that there are indeed journalistically relevant pieces of information being shared on Twitter. We manually categorized messages from the sample and identified contributions such as: quotes of the speech, observations from the scene including the environment and situated response, comments on the appearance of key actors, sentiment evaluations of the tone or content, interpretations of political implications, and intertextual ideological associations. These activities suggest that if properly connected to a visual analytics tool, journalists could harness the wisdom of the social media crowd to ultimately do better reporting of events.

We begin in the next section by describing the text analysis algorithms that support the journalistic goals and enhance the use of Vox Civitas as a visual analytics tool for journalistic inquiry.

4 Computational Enablers

Evaluations of visual analytics systems such as Jigsaw [13] have highlighted the importance of designing to jumpstart the analytical process by directing attention to relevant information and providing appropriate starting points for analysis. In order to facilitate these objectives in Vox Civitas we leverage four types of automatic content analysis: *relevance, uniqueness, sentiment,* and *keyword extraction*. These automatic analyses provide capabilities both for searching and filtering raw information in journalistically meaningful ways, as well as providing aggregate values (e.g. of sentiment) that can inform analyses.

4.1 Relevance

Assessing the relevance of social media messages is important for helping analysts reduce the amount of noise and focus on information more relevant to the event. We define relevance of social media messages with respect to the underlying event content: the transcript of the spoken words in the event. For many large-scale news events, such as the State of the Union, transcripts are readily available from news services. Our definition of relevance also incorporates a temporal component by assessing relevance for a message at a *particular* point in time. We acknowledge that different definitions of relevance could lead to different types of analytic capabilities.

We computed relevancy by calculating term-vector similarity of messages to the moment in the event during which the messages were posted. In order to compute relevancy at a finer level of granularity than the entire event, we further structured the raw transcript by breaking it into one-minute segments, and consider the text from each segment as the basis for relevance. For each message, relevance was computed as the cosine distance [16] of the term-vector space representations of the message and of the transcript for the minute when the message occurred (the transcript and messages were first filtered through a standard stop word list). To control for possible lag in the social media response, we used a running window (with weighting) over the previous two minutes. This method is designed to account for some delayed reaction to the speech, and compute a temporally sensitive relevance score, rather then assess the relevance of messages with a potentially unlimited lag. To calculate the relevance of a social media message at time m (S_m) to a particular minute m of the speech we use the transcript at time m (T_m) and associated term vectors,

$$rel(S_m, m) = 2 \times \frac{\vec{V}(S_m) \bullet \vec{V}(T_m)}{\left|\vec{V}(S_m)\right|\left|\vec{V}(T_m)\right|} + \frac{\vec{V}(S_m) \bullet \vec{V}(T_{m-1})}{\left|\vec{V}(S_m)\right|\left|\vec{V}(T_{m-1})\right|}$$

4.2 Uniqueness

Definitions of "newsworthiness" and "news values" in journalism often espouse the importance of the *unusual* or *unexpected* in the selection criteria for what becomes "news" [9, 11]. In the context of social media, "unusual" may manifest itself as more *unique*

messages when compared to other social media messages provided that the messages are still relevant to the event (see Figure 1). We incorporate this concept into our system by developing a message uniqueness metric, which can be used to direct attention toward what may be more unusual contributions.

Here we define the uniqueness of a message in relation to the other messages sent during the same time interval. It is computed as the difference between the term-vector space representation of the message to the centroid term-vector representation of all of the messages for that particular minute of the event (as above, messages were filtered through a standard stop word list). The centroid vector for each minute is constructed from the top 200 most frequent terms for that minute. For a social media message at time m (S_m) and the centroid for aggregate minute m (C_m) we calculated uniqueness as:

$$uniqueness(S_m) = 1 - \frac{\vec{V}(S_m) \bullet \vec{V}(C_m)}{\left|\vec{V}(S_m)\right\|\vec{V}(C_m)|}$$

A message that uses words unusual for that minute will not share many words with the centroid and will thus have a low cosine similarity score. We then define the "journalistically interesting" range of uniqueness for the filter presented in the interface by thresholding uniqueness values between a minimum and maximum value as suggested in Figure 1.

4.3 Sentiment

Sentiment analysis can be broadly construed as facilitating the understanding of opinion, emotion, and subjectivity in text [18]. Here we focus more specifically on sentiment analysis to inform an analyst's understanding of the *polarity* (i.e. positive versus negative) of the social media reaction to the event. Some of our prior work has shown that sentiment analysis of social media text polarity can inform analyses of the aggregate reactivity of the audience to an event topic, issue, or actor [7].

Classifying social media messages from sources such as Twitter poses a significant challenge. Despite considerable progress in the maturation and accuracy of sentiment polarity classification algorithms, these algorithms are still far from perfect [18]. Exacerbating the problem is the fact that social media content, often due to constraints on message length, is riddled with irregular language such as inconsistent abbreviations, internet-speak and other slang, and acronyms. Attempting to handle these issues, we followed a two-step procedure: we first ran a simple classifier based on a lexicon of words that classified messages based on whether they were carrying subjective (positive or negative) information [21]. In a second step we applied a supervised learning algorithm (language model) trained with 1900 manually tagged messages from the State of the Union corpus together with messages tagged as "other" by the simple classifier. We found the best performance using a language model including all n-grams of length less than or equal to four. The combined classifier resulted in a 5-fold cross validated accuracy of 62.4%. This is sufficient for giving an overall impression of the sentiment, but the classifier still fails on difficult cases involving sarcasm or slang. For example, "*whats goodie twiggaz..im watchin Obama talk about how he gna clear my student loans..i kno there was a reason i voted for him lol!*" was classified as negative by the algorithm but is arguably positive.

4.4 Keyword Extraction

In keeping with the design goal of jump-starting analysis, we aimed to identify keywords used in the social media stream that could be useful and interesting for guiding analysts. To this end, we extracted descriptive keywords for each minute of the aggregate message content. For each minute we extract the top 10 keywords ranked by their tf-idf score [16], comparing the keyword's frequency at that minute to its frequency in the rest of the dataset. We found tf-idf performed adequately for identifying salient keywords, although other methods for extracting salient key phrases [22] or words [4] could be implemented and integrated into our data processing pipeline. For the purposes of the document frequency in our tf-idf scores we define pseudo-documents temporally as the aggregate of the words of all messages for each minute. Words are first stemmed using the Porter stemming algorithm and after computing tf-idf scores on word stems we apply reverse stemming to the most common full keyword mapping so that complete words are visible in the interface [4].

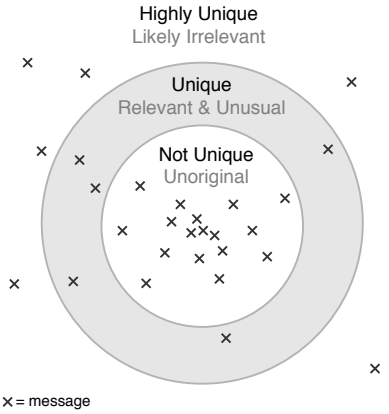

Figure 1. A conceptual diagram showing the relationship between message utility in the context of journalism with levels of uniqueness. Those messages in the middle gray band may be newsworthy in the sense that they are unusual but still relevant to the discussion.

5 Visual Representations and Interactions

The Vox Civitas interface integrates video from an event with the ability to visually assess the textual social media response to that event at both (1) individual, and (2) aggregate levels of analysis. The unifying schema for organizing information in Vox Civitas is temporal, which facilitates looking at responses and trends over time in the social media stream, in relationship to the underlying event video. Figure 2 shows an overview of the interface.

Filtering messages is done via the module shown in Figure 2A. Browsing and analysis of individual responses is facilitated by a view of the actual Twitter messages posted about the event (next to the video content, in Figure 2B). Aggregate response analysis is enabled by three views: volume graph (2E), sentiment timeline (2F); and the keywords component (2G). These views are all aligned to the video timeline (2C) and the topic timeline (2D) and are connected visually to the timeline via a light gray vertical bar which tracks the navigation thumb of the video timeline. In the rest of this section, we explain the main interactive elements of our interface. For each element, we explain the interaction and, where appropriate, how the interaction builds on the computational foundations laid out above.

5.1 Content Component

The content component (Figure 2B) displays the "raw" content from the event and its social media response. On the left, the video feed from the event is shown. The video is controlled by the timeline (Figure 2C) that allows start, pause and nonlinear navigation of the video steam. On the right side of Figure 2B, the interface shows the set of messages about the event that were

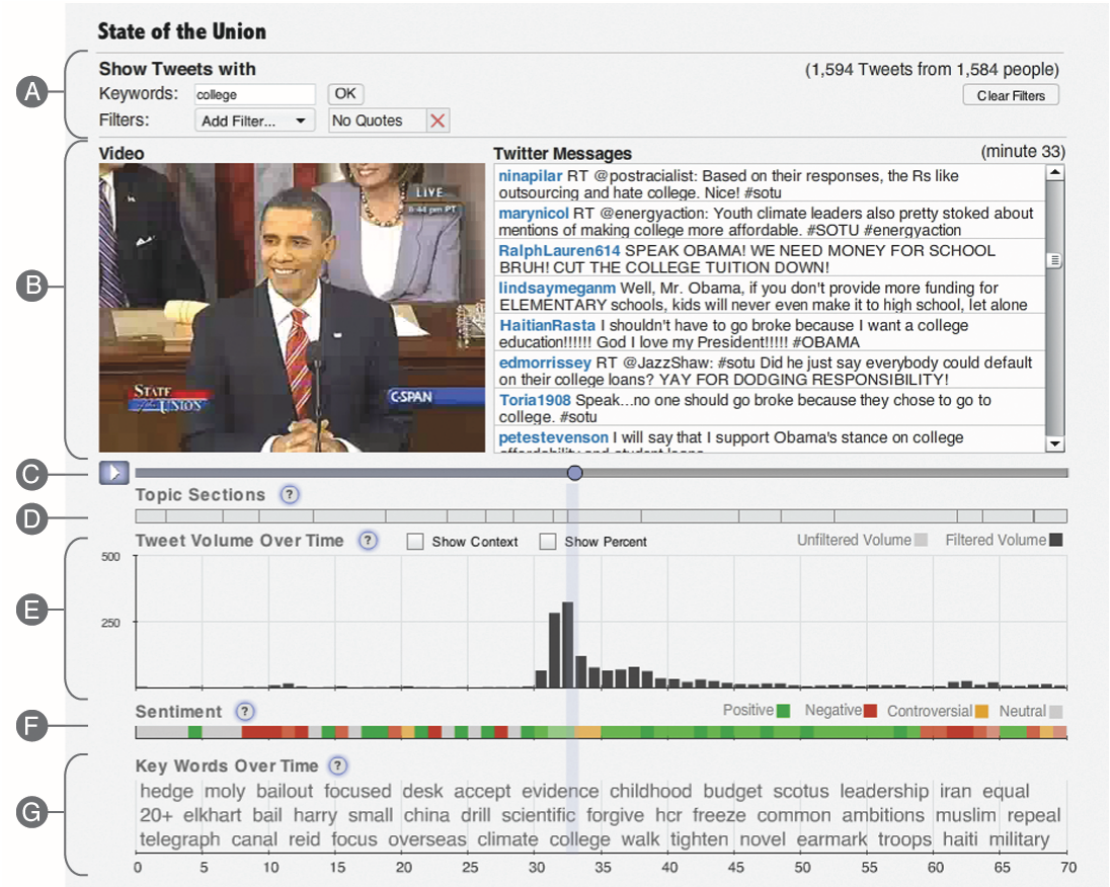

Figure 2. Vox Civitas User Interface. (A. Keyword Search & Filtering, B. Video & Twitter Messages, C. Video Timeline, D. Topic Sections, E. Message Volume Graph, F. Trends in overall Tweet Sentiment, G. Salient Key Words Over Time)

posted during the minute currently selected by the user via the various timeline interactions, mirroring the currently-viewed portion of the video. The messages displayed can be filtered using the filtering module as we explain next.

5.2 Filtering Module

The filtering module, shown in Figure 2A, allows Vox Civitas users to filter the social media responses to the event according to a number of criteria or search terms. The filtering has a number of outcomes: it determines which messages are displayed in the main content pane (2B), as well as the aggregate statistics in the message volume graph (2E) and sentiment timeline (2F). All these components update interactively with the filter. In our current implementation, filtering does not change the keyword pane (2G).

The filtering module options build on some of the computational aspects described above. Users can apply the following filters to the messages, individually or in conjunctive combination: (1) messages with specific keywords or authors, (2) messages with quotes (i.e, quotation marks), (3) messages that are retweets (messages that are repeated or forwarded from other users and usually marked "RT" at the beginning), (4) messages that include links, (5) messages classified as being topically relevant (Section 4.1 above), (6) messages classified as unique (Section 4.2 above), and (7) messages with positive or negative sentiment (Section 4.3 above). The system also allows for filtering out quotes, retweets or links. In Figure 2, for example, the active filters are the keyword "college" and "no quotes".

5.3 Topic Timeline

We included a topic segmentation timeline (Figure 2D) that facilitates building connections between topicality, time, and the social media response. The topic timeline shows the temporal extent of topic sections of the speech and is aligned with the video timeline (Figure 2C). Hovering over a topic section shows the section's label and clicking navigates the content component (video and messages) to the beginning of that section. Note that the topics appearing on the timeline and their time range could be automatically detected from the message or the event content, or provided by a human editor.

5.4 Message Volume Graph

The message volume graph (Figure 2E) shows the message volume over time as a histogram, where each bar represents one minute. The heights of the bars represent the aggregate volume of messages according to the currently applied filter. By default, the overall volume of messages is shown. Changes to the filters initiate an animated transition on the graph so that differences can be tracked visually. Check boxes allow the user to compare the current filtered set to the total volume, as well as change the vertical scale from absolute to percent in order to assess the proportional filtered response over time. Hovering over the graph shows a popup of the exact count or percent of messages at that minute as well as the number of unique users contributing to those messages. The message volume graph also acts as an interactive timeline: clicking the graph navigates the video and messages of the content component to that minute.

5.5 Sentiment Timeline

The sentiment timeline (Figure 2F) shows an aggregate of the sentiment response for each minute of the event, as derived by the sentiment analysis described in Section 4.3. The timeline is color-coded according to one of four categories: positive (green), negative (red), controversial (yellow), or neutral (gray). A minute is categorized as controversial if the ratio of positive to negative messages for that minute is between 0.45 and 0.55. If there are no positive or negative messages then that minute is categorized as neutral. If either positive or negative messages dominate the dataset for that minute (the ratio is above 0.55) then that minute is categorized as positive or negative respectively. The coloring for positive or negative minutes has five grades of intensity depending on the ratio of how much one sentiment dominates the other. Hovering over the sentiment graph will show the detailed counts of positive and negative classifications of messages for that minute. The sentiment timeline changes to reflect the currently applied filter, and is interactive: clicking navigates the video and messages of the content component to that minute.

The sentiment representation is explicitly designed to give only an *impression* of *aggregate* sentiment due to concerns over the accuracy of the sentiment classifier. We do not represent the automatic sentiment classification of individual messages in the message list (Figure 2B) since we assume users can quickly surmise sentiment as they are skimming the short text messages. Also, we do not represent absolute magnitude of the aggregate sentiment response (or show the distribution of positive and negative magnitudes). Journalists that we spoke to early in the design process believed that until the accuracy of the classifier was ~70-80% or higher, visual representations could easily mislead the analyst if they showed absolute magnitudes. Our visual representation helps cope with the depiction of uncertainty in the accuracy of the sentiment classifier by not giving undue weight to the comparative magnitude of positive versus negative messages. Moreover, if we assume that the error in the classifier is uniformly distributed in time, temporal sentiment trends are still meaningful.

5.6 Keywords Component

The keywords component (Figure 2G) depicts salient keywords over time, extracted as described in Section 4.4. It is similar to a "tag cloud" that has been laid out so that word positions are correlated with the time span when the chosen word was most salient in the event. We chose to keep the visual depiction simple by not visually encoding any additional facets of information (e.g. degree of salience into color intensity or font size) beyond just the keyword and its approximate time-span. Clicking on a word in the keyword component filters the dataset using that keyword, which in turn affects the other components as described above.

The component is laid out from left to right and top to bottom using a greedy algorithm. For each minute, we have a list of salient keywords ranked by their tf-idf scores. For a given layout position we compute the layout score of a proposed keyword as the sum of the word's tf-idf scores for all minute intervals that the keyword would span when laid out. So for example, if a word when added to the component would span 5 minutes worth of space, that word's score is the sum of its tf-idf scores for all of those 5 minutes. This way, we give preference to words that are potentially relevant for more than a single minute in time. For each time position, we select the keyword with the highest layout score, add it to the layout, and advance to the next position (after the current word plus a padding offset). Once a word has been added to the component it is removed from the ranked lists of keywords for the minute intervals it spans. This prevents duplicate words being added to the component adjacent to each other, but also allows duplicate words if they are relevant at different sections of the event. The depth of the layout can be expanded to include as many rows of words as desired.

6 Exploratory Study

We designed and executed an exploratory evaluation of Vox Civitas to assess its effectiveness in a journalistic reporting scenario performed by the application's target audience, namely journalists and media professionals. The goals of the evaluation were to develop an understanding of how journalists use the tool, and how Vox Civitas matches the journalists' requirements and work process. We address these research questions:

- How useful and effective was the tool for journalists in generating story ideas and reporting on the event?
- What kind of insights and analysis does Vox Civitas support?
- What are the shortcomings of Vox Civitas for journalists analyzing social media streams?
- How do journalists interact with the system and which parts of the interaction are most salient?

To answer these questions, we deployed Vox Civitas using a popular broadcast event as a content source, recruited participants with a background in journalism, and deployed the system while collecting questionnaire feedback and analyzing interaction logs. We analyzed the open-ended questionnaire items using a grounded-theory inspired methodology. This methodology involves iterative coding of concepts and their relationships apparent in the text in order to form typologies of use and patterns of interaction grounded in the participants' textual response data.

6.1 Content

We structured the evaluation of Vox Civitas around the State of the Union address by U.S. President Barack Obama in early 2010. This broadcast event is traditionally heavily covered by media, and generates high news interest. We anticipated the event would result in a large social media response on Twitter and other forums. Indeed, immediately after the event we collected 101,285 English language Twitter messages containing the terms "SOTU" (for "State of the Union"), "Obama", or "State of the Union" using the Twitter API. This keyword-based sampling method does not ensure collection of *all* relevant messages for the event (relevant messages not containing these terms will not be retrieved). However, we believe that the resultant dataset is more than adequate for enabling the sensemaking capabilities of the interface.

Once retrieved, we analyzed the 101,285 messages to detect relevance, uniqueness, and sentiment (Section 4). In total, the algorithms marked 15,312 messages (15% of all messages) as "relevant", 12,110 messages (12%) as "unique", 24,487 messages (24%) as positive, and 54,043 messages (53%) as negative. We used an upper threshold of 0.99 and a lower threshold of 0.95 for uniqueness, and a threshold of 0.3 for relevance to obtain those numbers. These thresholds seem to work in practice, but we leave it for future work to optimize and further evaluate these values.

6.2 Procedure

Vox Civitas is a Web-based system[1] and the evaluation was conducted online. We chose an online evaluation rather than a lab study to enhance the ecological and external validity of the study. The experiment was deployed using "natural" settings in terms of work environment, time constraints and so forth. The online nature of the deployment also enhanced the ability to include a larger number of journalism professionals from around the nation. We logged the participants' actions with the interface and recorded open-ended survey responses. We identified interactions

[1] http://sm.rutgers.edu/voxcivitas/voxcivitas.html

or survey responses too short to be meaningful and excluded one response from our analysis as a result.

To solicit participation, a convenience sample of journalists and journalism students was emailed with a request to participate in our study, for which they were entered into a drawing to win a $50 gift card. The call for participation was also published in other venues that we thought likely to bring participants (e.g., Twitter and mailing lists). Participants were directed to a website, where, upon consent to become a research participant, they were presented with the Vox Civitas system. An overview description of the tool and its functionality was displayed next to the tool itself, briefly explaining to the users the interface's main features.

The instructions and scenario for the task were persistently displayed next to the tool. The participants were instructed to act as journalists performing a task, namely using the tool to find stories to pitch to a national news editor, shortly after the State of the Union address took place. The participants were then asked to develop at least two story angles, which they thought would make good stories. Their interactions with the interface (mouse hovers, clicks, time spent using the tool) were logged.

When ready with their story angles, participants were asked to fill out an online questionnaire. The bulk of the questionnaire was composed of open-ended questions including the story angles the participants developed, the ways in which the tool enabled them to develop the story angles, how they would use such a tool to inform their reporting on a broadcast media event, and what they liked or disliked about the user interface. The questionnaire also included demographic data, as well as questions about the participant's training in journalism and the frequency of their usage of social media services.

6.3 Participants

Eighteen participants were recruited, 15 of which had formal or on the job training in journalism according to their responses: seven participants identified themselves as professional journalists, five as journalism students, and one as a citizen journalist; two additional participants did not identify as journalists but specified that they had an undergraduate degree or "on the job" experience in journalism. Six respondents were male, and twelve were female. The ages of the participants ranged from 21 to 55 (μ=34). Eleven of the participants indicated that they use social media services such as Twitter all the time, while five indicated they use them "often", and only two indicated that they use them "sometimes".

6.4 Results

We first report on our findings based on the grounded analysis of open-ended questionnaire items. We then briefly report on the usage of the application and its various features as captured by the interaction log.

6.4.1 Perceived Utility

In the questionnaire, participants answered the open-ended question "*If you were to use this or a similar application to inform your reporting on a broadcast media event, how would you use it?*" We used a grounded approach to categorize and code the open-ended responses to this question. We identified two primary use cases for Vox Civitas: (1) as a mechanism for finding sources to interview and (2) as an ideation tool for driving follow-up journalistic activity.

Finding and interviewing credible primary sources is an important aspect of journalistic storytelling [14]. Indeed, prior studies assessing tools for journalists have shown the primacy of sourcing in appealing to the journalistic mindset [6]. As such, it is perhaps unsurprising that several users of Vox Civitas suggested it would be a valuable tool for helping to identify sources. As one participant put it, "*I might use it to track sources reacting to an event that I could quickly turn to for an interview*" (P11).

Beyond sourcing, several participants noted that Vox Civitas would be useful for helping to find unusual story angles and statements that resonated with the audience: "*I would use it for drilling down to the outlier sentiments in response to the State of the Union*" (P16) and "*Using the quotes and retweet filters, I can also easily figure out what statement resonated the most with the public"* (P8). These responses reaffirm the newsworthiness values which Vox Civitas was designed to support, such as helping to identify unique contributions, or "decisive moments" which draw heavy audience response [3, 11]. Other participants identified related uses for driving journalistic activities such as helping to *measure interest* for particular follow-up stories or as an input to a discussion panel after the event.

It is important to note that, while predominantly positive in their outlook for Vox Civitas, several responses indicated healthy suspicions about relying solely on the tool for reporting. Concerns revolved around the recognition that Twitter does not represent an accurate population sample for measuring global sentiment, and that tools like Vox Civitas are useful "*as long as they are used as a compliment to stories that include more sound data*" (P4) or "*as a jumping off point for stories, as long as it is clear that the tweets aren't representative for the whole country*" (P9).

6.4.2 Story Angles

In order to assess more specifically what kinds of story ideas and types of insight might be generated with Vox Civitas we asked participants to consider the scenario of using the tool to come up with two story angles they might pitch to a news editor of a national publication. Of course, this scenario serves only as a (reasonable) proxy for real journalistic practice, since real story angles would depend on the context of publication and audience.

Again, we used a grounded approach to categorize and iteratively code participants' open-ended responses. We address the types of story angles, as well as the Vox Civitas features that drove and enabled the development of stories.

Two main foci of story angles emerged from our analysis: stories that focus and reference the *event content* and stories that reference *audience responses* to the event. Event content story angles focused on topics, issues, or personalities in the event such as words spoken, or the body language or appearance of people in the video. Most often, these story angles referred to topics or issues that were referenced in the speech. One story pitch read:

> *"Obama's plan to increase Pell grants: What kind of students, majors and schools would give the government the best return on their investment to get the kinds of workers the country needs?" (P1)*

Notice that these stories emerged from examining Vox Civitas, but the participants did not *directly* reference the social media response in their story angle.

On the other hand, story angles referencing *audience responses* focused on the reactions of the audience to the event, including both reactions captured in individual messages, as well as the magnitude or sentiment of the aggregate audience response to various aspects of the event and the issues being discussed therein. One participant wrote:

> *"The two topics that did create a 'controversial' exchange were 'People's Struggles' and 'Stimulus: Tax cuts and Employment.' This could compliment other data about job losses and the economy and … could make for*

an interesting angle on what topics in public sentiment are most polarized." (P4)

A more minor focus (two story angles in our survey) was *audience meta discussion*. These stories focused on the characteristics of the audience in the social media channel (e.g., its demographic), rather than the audience response to the event.

How exactly did Vox Civitas support the creation of these story ideas? Some participants reported that their story angles were informed through the use of keyword searches and further filtering (e.g. sentiment) to help them identify individual or aggregate responses. One participant started an inquiry in response to an individual tweet she saw referencing low college loan payments. The story angle read: "*Further investigation into statistics on college loan debt. How much are students carrying? And how long does it take to pay off?*" *(P16)*. Other participants, like P4 above, looked to aggregate cues such as the magnitude or sentiment of a response to a keyword or topic to drive ideation:

"I liked using the keywords to elicit the popularity of a certain topic. For example, 'college' was probably by far the most powerful statement, showing 500 tweets immediately after Obama's 'no one should go broke' statement…" (P8).

"I chose the keyword 'overseas'. This gave me more of mixed emotions for the audience due to the slash in tax breaks being given to companies who ship their jobs overseas" (P18).

Indeed, many of the story angles that were reported mentioned aspects of the visuals and interface that were used to enable those thoughts. We turn briefly in the following section to aspects of the log analysis that further support these findings.

6.4.3 Usage of Interface Features

We see the results of the log analysis as illustrative and use them to support our findings on the utility of features for journalists, although we did not have enough participants to be able to derive statistically meaningful patterns from the log data. Participants spent an average of 21.6 minutes interacting with the application, with 89% of users spending more than five minutes.

The utility and popularity of searching for keywords and combining those searches with further filters was evident in the logs. All 18 participants performed some keyword searching and filtering activity. Users searched for an average of 9.67 unique words each (σ=10.2). Half of the participants also used compound filters, meaning they combined a keyword search with a filter modifier. Among these, two people filtered for relevancy, three for uniqueness, six for negative sentiment, two for positive sentiment, two for retweets, four for no retweets, three for quotes, and one each for no quotes and links. Judging from these counts, filters for sentiment and retweets were used most in conjunction with the keyword filters, with other filters used to a lesser extent.

An average of 4.67 keyword searches per user were initiated from the keywords over time component, meaning that 48% of all keyword queries came from users interacting with that interface feature (the remaining keyword queries were initiated by users typing words into the search box). However, we note that only eight of the 18 participants clicked to filter by a keyword via the keyword component, with five users making heavy use of the component to drive the filtering. When we looked at the use of the keyword component by professional journalists versus all others (students, citizen journalists, and non-journalists) there was a clear trend of the professionals using the keyword component *less*: only one journalist used it.

The topic timeline, volume graph, and sentiment timeline all saw robust usage in terms of users gleaning data details from hovering over these representations. Sixteen out of 18 users hovered over the topic timeline (mean of 34 operations per user) and when normalized for interaction duration, seven users averaged more than one hover operation per minute of use. A total of 17 users hovered over the volume graph (μ=392) with 15 users averaging more than one hover operation per minute. Similarly, 15 users hovered over the sentiment timeline for details (μ=54) and 13 users averaged more than one sentiment hover operation per minute. Combined with the prevalence for searching and filtering for keywords, these numbers tend to indicate that users informed their analyses by employing the volume graph most, followed by the sentiment timeline, and topic timeline.

7 Discussion

Our results suggest that Vox Civitas' utility is in divergent modes of sensemaking, where the tool is used to (1) drive analysts to gather information from identified sources, and (2) to otherwise inform journalists in more "creative" follow-up activities such as finding unusual story angles, or as a *starting point* for further inquiry on a topic or sentiment reaction. The journalistic goal in this use case is not so much to provide rigorous assessment and decision support about hypotheses, but rather to spur the divergent and creative generation of hypotheses, insights, and questions for follow-up activities.

Let us consider a sensemaking model such as that of Pirolli and Card [20], which consists of *information foraging* (collecting from external data sources, shoeboxing, building an evidence file) and *sensemaking loops* (scheme generation, hypothesis generation, and final presentation). Vox Civitas seems to best support aspects of hypothesis generation in the sensemaking loop, as well as rapid transition back to the foraging loop in terms of facilitating connecting to external data sources. This support was assisted by the journalistically-motivated design that provided for a sensemaking schema, thus organizing the information visually according to cues expected to be of interest (topic, magnitude, sentiment, uniqueness) to the target audience.

Vox Civitas obviates the initial phases of the sensemaking process (data collection and schema generation) and allows analysts to "skip" to divergent thinking and hypothesis generation around the data. This divergent thinking can then connect back to the foraging loop to collect data from external sources to support a follow-up story. We believe that designers of similar visual analytics systems may be able to extend this notion to other domains of expert analysts by tailoring filtering and initial visual scheme presentation in order to jump start the sensemaking process at a high level of thinking.

The keywords component drove a substantial portion of the keyword searching and filtering activity, albeit the utility of this component for professional journalists may be less than for citizen journalists or student journalists. Nonetheless, the component raises the idea of driving different people to different parts of the information space so as to jumpstart analysis along different dimensions. For instance, we could imagine producing a keyword timeline that varies depending on the news genre that someone is interested in reporting on. Keywords for business, sports, technology, or fashion would tend to drive analysts to think about those term-sets in relation to the event.

Amar and Stasko's [1] first suggestion for dealing with the rationale gap (the gap "between perceiving a relationship and actually being able to explain confidence in that relationship") is to *expose uncertainty*. In the design of our sentiment visualization timeline we were forced to accept the limitations of the accuracy of our automatic classifier and in doing so approached the visualization of uncertainty a bit differently: we reduced the precision of the visual representation commensurate with the degree of uncertainty. More specifically, we chose to deal with the

depiction of uncertainty in the sentiment classifier by not giving undue weight to the representation of comparative magnitude between positive and negative aggregates. Of course, the user is left with less information (and no accurate depiction of the uncertainty in the classifier), but the interface is simpler, and users are not lead to assume relationships that are inaccurate.

The evaluation of information visualization and visual analytics systems has been acknowledged as one of the defining challenges of the field. The approach we have taken, an online evaluation which links concrete but open-ended analytic insights to user's interactions with the interface, is a promising evaluation methodology, which if scaled up to include more users would lead to an ability to run statistical tests between interaction patterns and coded analytic outputs produced in ecologically valid situations.

8 Conclusion

Journalists turn to social media for story angles, leads and even to obtain rough (yet immediate) proxies of public response. We presented a tool to support media professionals in achieving these goals by collecting, analyzing, aggregating and visualizing content from one major broadcast event, the US presidential State of the Union address of 2010. We have shown that journalists (and others) effectively use the tool to generate insight about the social media response to the event, and about the event itself.

In future work, we intend to generalize the application and verify its utility for different types of broadcast events, from entertainment (e.g., The Academy Awards) to televised breaking news. We also intend to enhance and improve the automated analysis tools, and in particular the sentiment analysis, which can now only provide very general trends. We plan to extend the reach and usefulness of the application by allowing users to illustrate a point by embedding a selected visualization state in any webpage.

Finally, we believe there is significant space for computational and technical innovation to more directly support core journalistic tasks while adhering to the journalistic values of accuracy, objectivity, and impartiality. For example, potential advances can leverage network structure as well as content and activity volume to extract metrics of expertise. In combination with sentiment analysis, such metrics could help characterize sources in terms of bias, and in turn help inform a journalist's selection of sources to better balance the reporting on a story. Developing such tools will be key for future systems that report on all the news that's fit to tweet.

9 Acknowledgments

Financial support for this work was provided through the CRA and NSF as part of a Computing Innovation Fellowship (CIF-197). Thanks to Hrishikesh Bakshi for help with dataset collection. We would also like to thank Smaranda Muresan, Susan Keith, and Steve Miller as well as journalists from the New York Times and American Public Media for feedback on early versions of our system interface and functionality.

References

[1] Amar, R.A. and Stasko, J.T. Knowledge precepts for design and evaluation of information visualizations. Visualization and Computer Graphics, IEEE Transactions on, 11 (4). 432-442.

[2] Chin Jr., G., Kuchar, O.A. and Wolf, K.E., Exploring the Analytical Processes of Intelligence Analysts. in Proceedings of CHI, (2009).

[3] Clayman, S. Defining Moments, Presidential Debates, and the Dynamics of Quotability. Journal of Communication, 45 (3).

[4] Collins, C., Viégas, F. and Wattenberg, M., Parallel Tag Coulds to Explore and Analyze Faceted Text Corpora. in IEEE Symposium on Visual Analytics Science and Technology (VAST), (2009).

[5] De Longueville, B., Smith, R. and Luraschi, G., "OMG, from here, I can see the flames!": A Use Case of Mining Location Based Social Networks to Acquire Spatio-temporal Data on Forest Fires. in Workshop on Location Based Social Networks (LBSN), (2009).

[6] Diakopoulos, N., Goldenberg, S. and Essa, I., Videolyzer: Quality Analysis of Online Informational Video for Bloggers and Journalists. in Proceedings of CHI, (2009).

[7] Diakopoulos, N. and Shamma, D.A., Characterizing Debate Performance via Aggregated Twitter Sentiment. in Proceedings of CHI, (2010).

[8] Fisher, D., Hoff, A., Robertson, G. and Hurst, M., Narratives: A Visualization to Track Narrative Events as they Develop. in IEEE Symposium on Visual Analytics Science and Technology (VAST), (2008).

[9] Franklin, B., Hamer, M., Hanna, M., Kinsey, M. and Richardson, J.E. Key Concepts in Journalism Studies. Sage Publications, 2005.

[10] Gregory, M., Chinchor, N., Whitney, P., Carter, R., Hetzler, E. and Turner, A., User-directed Sentiment Analysis: Visualizing the Affective Content of Documents. in Proceedings of the Workshop on Sentiment and Subjectivity in Text, (2006).

[11] Harcup, T. and O'Neill, D. What is News? Galtung and Ruge Revisited. Journalism Studies, 2 (2). 261-280.

[12] Havre, S., Hetzler, E., Whitney, P. and Nowell, L. ThemeRiver: Visualizing Thematic Changes in Large Document Collections. IEEE Transactions on Visualization and Computer Graphics, 8 (1).

[13] Kang, Y.-a., Görg, C. and Stasko, J., Evaluating Visual Analytics Systems for Investigative Analysis: Deriving Design Principles from a Case Study. in IEEE Symposium on Visual Analytics Science and Technology, (2009).

[14] Kovach, B. and Rosenstiel, T. The Elements of Journalism: What Newspeople Should Know and the Public Should Expect. Three Rivers Press, 2007.

[15] Leskovec, J., Backstrom, L. and Kleinberg, J., Meme-tracking and the Dynamics of the News Cycle. in Conference on Knowledge Discovery and Data Mining (KDD), (2009).

[16] Manning, C., Raghavan, P. and Schütze, H. Introduction to Information Retrieval. Cambridge University Press, 2008.

[17] Nagar, N.A. The Loud Public: Users' Comments and the Online News Media. Online Journalism Symposium, 2009.

[18] Pang, B. and Lee, L. Opinion Mining and Sentiment Analysis, 2008.

[19] Pioch, N. and Everett, J., POLESTAR: Collaborative Knowledge Management and Sensemaking Tools for Intelligence Analysts. in Conference Information and Knowledge Management, (2006).

[20] Pirolli, P. and Schneiderman, B., The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. in International Conference on Intelligence Analysis, (2005).

[21] Riloff, E. and Wiebe, J., Learning Extraction Patterns for Subjective Expressions. in Empirical Methods in Natural Language Processing (EMNLP), (2003).

[22] Rose, S., Butner, S., Cowley, W., Gregory, M. and Walker, J., Describing Story Evolution from Dynamic Information Streams. in IEEE Symposium on Visual Analytics Science and Technology (VAST), (2009).

[23] Schudson, M. The Sociology of News (Contemporary Sociology). W. W. Norton & Company, 2003.

[24] Shamma, D., Kennedy, L. and Churchill, E., Conversational Shadows: Describing Live Media Events Using Short Messages. in Proceedings of ICWSM, (2010).

[25] Shamma, D.A., Kennedy, L. and Churchill, E. Tweet the debates ACM Multimedia Workshop on Social Media (WSM), 2009.

[26] Starbird, K., Palen, L., Hughes, A. and Vieweg, S., Chatter on The Red: What Hazards Threat Reveals about the Social Life of Microblogged Information. in Proceedings of CSCW, (2010).

[27] Wanner, F., Rohrdantz, C., Mansmann, F., Oelke, D. and Keim, D., Visual Sentiment Analysis of RSS News Feeds Featuring the US Presidential Election in 2008. in Workshop on Visual Interfaces to the Social and Semantic Web, (2009).