Social Media Visual Analytics for Events

Nicholas Diakopoulos, Mor Naaman,
Tayebeh Yazdani, Funda Kivran-Swaine

Rutgers University, School of Communication and Information

4 Huntington St., New Brunswick, NJ 08901, USA

Abstract. For large-scale multimedia events such as televised debates and speeches, the amount of content on social media channels such as Facebook or Twitter can easily become overwhelming, yet still contain information that may aid and augment understanding of the multimedia content via individual social media items, or aggregate information from the crowd's response. In this work we discuss this opportunity in the context of a social media visual analytic tool, Vox Civitas, designed to help journalists, media professionals, or other researchers make sense of large-scale aggregations of social media content around multimedia broadcast events. We discuss the design of the tool, present and evaluate the text analysis techniques used to enable the presentation, and detail the visual and inter-action design. We provide an exploratory evaluation based on a user study in which journalists interacted with the system to analyze and report on a dataset of over one hundred thousand Twitter messages collected during the broadcast of the U.S. State of the Union presidential address in 2010.

Introduction

Social media systems have proven to be valuable sources for information and communication about media events, in particular during large-scale televised events. Events such as the annual Academy Awards (Oscar's) ceremony in Holy-wood, the Superbowl (the American National Football League's championship game), the last episode of a TV series, or other events such as televised speeches or emergency events increasingly draw massive amounts of audience attention and commentary via social media. This rush of information from millions of new "human sensors" contributing information about events and other news stories suggests new opportunities for reasoning both about the content of the event as well as about the crowd's response to it. However, the sheer scale of the contribution leads to the challenge of making sense of the response, both at an individual and aggregate level of analysis.

Social media content such as Twitter or Facebook posts offers real-time commentary about many multimedia events that can be used to reason about the event content. The stream of social media messages parallels and reflects on the multimedia content as it's being broadcast. This new stream – a crowd-sourced, interpretive "close-captioning" analogue – presents new opportunities for reasoning about the presented content using these text-based comments, which essentially serve as indirect annotations. The main assumption in this work is that most of the social media content posted in response to a broadcast event provides some reflection on a *specific segment* of the event, and that the segment's time is close to the time of the posting of the message.

In this work, we focus on how social media content contributed around large-scale broadcast news events can inform visual analytic investigation and produce insight regarding the response to the event. Here we define visual analytics as the use of a visual interface to amplify analytic reasoning, including both deriving insights from data as well as producing and communicating judgments based on that analysis [27]. Our use case and design process is motivated by the context of journalistic investigation. Journalists are increasingly turning to social media sources like Twitter, Facebook, and 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]. Here we detail the design and evaluation of a visual analytics system, Vox Civitas, whose goal is to make the social media (e.g., Twitter) response to broadcast more amenable to journalistic investigation and sensemaking. Furthermore, we report on our evaluations of the various text analysis components that enable the Vox Civitas interface, such as relevance, novelty, and sentiment detection algorithms.

Related Work

Our work is most inspired by the work of Shamma et al. [23, 24] 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 analytic purposes.

Other related work has examined social media content as an information source for non-broadcast events, in the context of emergency response and crisis scenarios such as earthquakes [22], 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 ac-

tivities 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.

The analysis of text corpora over time has been addressed by a variety of systems including ThemeRiver [11], 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 [14] 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* [21]. 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 analysis of text corpora derived from social media communication in order to better reason about multimedia content has been proposed in a number of prior studies. Most directly related to our work here, Diakopoulos and Shamma [7] presented temporal visuals which depict sentiment patterns (e.g. periodicity, strength or weakness of actors) in the context of the real-time social media response to the televised U.S. presidential debates. Shamma et al. [25] use chat activity in instant messaging to reason about the content of shared online videos; De Choudhury et al. [4] analyze comments on YouTube videos to derive interestingness and topics; and Mertens et al. use community activity for social navigation of web lectures [16]. Here, we go beyond these prior systems to connect text patterns with topicality and the magnitude of the response in a visual and interactive exploratory analytics application.

Computational Enablers

Previous work on the evaluation of visual analytics systems has highlighted the importance of directing attention and providing appropriate starting points for analysis [12]. In this section we begin by describing each of the content analysis components of Vox Civitas that enable different aspects of filtering and attention shaping in the resulting interface. We leverage four types of automatic content analysis to do this: *relevance, novelty, sentiment,* and *keyword extraction*. These automatic analyses provide capabilities both for searching and filtering raw information in analytically meaningful ways, as well as providing aggregate or derivative values (e.g. of sentiment) that can inform analyses.

Where appropriate we validate and characterize our computational techniques and parameters through comparisons with ground truth human ratings. For these evaluations as well as for the interface evaluation (presented in a later section) we gathered a collection of Twitter messages from the State of the Union address given by U.S. President Barack Obama in early 2010. This broadcast event is tra-

ditionally heavily covered by mainstream 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. Note that 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.

Relevance

Assessing the relevance of social media messages is important for helping to reduce the amount of noise and focus on information more germane to the event. We define relevance of social media messages with respect to the underlying audio channel of event content: specifically, the topics expressed in the transcript of the spoken word audio channel of the event. For many large-scale televised events, such as the State of the Union, transcripts are readily available from news services such as C-SPAN. 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 can lead to different types of analytic capabilities and we explore this idea further in our evaluation below.

We compute relevance by calculating term-vector similarity of messages to the moment in the event during which the messages were posted. In order to compute relevance 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. We chose a segmentation interval of one minute because (1) it is a meaningful unit of analysis for events on the scale that we are currently concerned with (e.g. a 70 minute speech), and (2) one minute of transcript provided enough text to create meaningful term vectors.

For each message, relevance was computed as the cosine distance [15] 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. This enables filtering for the *real-time* conversation happening on social media, and serves to dampen conversational echoes and shadows [23]. To calculate the relevance of a social media message at time m (S_m) to the transcript at time m (T_m) we take the time-interval weighted sum of the cosine similarity of the associated term vectors (see Fig 1),

$$rel(S_m, m) = 2 \times \frac{\vec{V}(S_m) \bullet \vec{V}(T_m)}{\left|\vec{V}(S_m)\right|\left|\vec{V}(T_m)\right|} + \frac{\vec{V}(S_m) \bullet \vec{V}(T_{m-1})}{\left|\vec{V}(S_m)\right|\left|\vec{V}(T_{m-1})\right|}$$
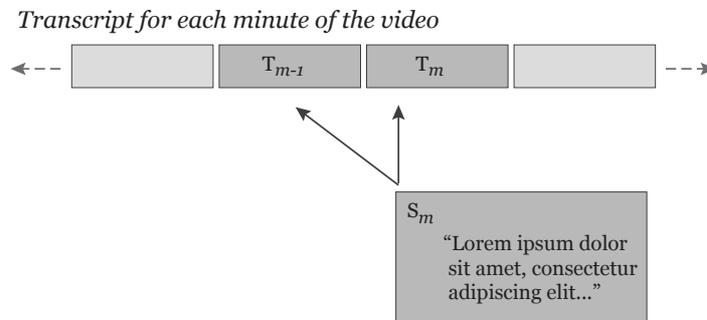
Transcript for each minute of the video

Fig 1 Relevance is computed for a social media message at time m by comparing to the text in the transcript at times m and m-1

In order to characterize and optimize how this operationalization of relevance affects the performance of the filter in Vox Civitas we undertook an evaluation to compare the automatically computed relevance scores to a manually coded ground truth. The ground truth was developed by first randomly sampling 2,000 messages from our State of the Union dataset. For each of these messages, one of the researchers then manually applied our definition of relevancy to classify each of the 2,000 messages as either relevant or not relevant (i.e. a binary classification). The definition of relevance used for the manual coding was that the message had to have some topical relevance to the transcript of the speech during that minute or the previous minute. For example, "*Hope Obama will show leadership on climate*" which was sent during the part of the speech where Obama was addressing climate change was tagged as relevant. In keeping with the temporal emphasis of our definition, general opinions (e.g. "Obama Rocks!") or status updates (e.g. "watching the state of the union") which were not germane to the content of the speech during that time period were classified as not relevant.

Applying our definition of relevance, we found that of the 2,000 ground truth messages, 29.6% of them were deemed relevant. Fifty-seven messages were relevant to the current minute without being relevant to the previous minute and 351 messages were relevant to the previous minute without being relevant to the current minute, with 185 messages being relevant to both minutes. These descriptive numbers indicate that there is some (expected, perhaps) lag in the relevancy of the messages with respect to the multimedia content (i.e. a majority of people need a bit of time to think, type, and respond appropriately).

The accuracy of the relevancy computation with respect to the ground truth necessarily depends on the threshold chosen to make the binary distinction between relevant or not relevant. We computed standard information retrieval measures such as precision, recall, and F-measure to better characterize the perform-

ance of the algorithm and show the tradeoffs associated with choosing a threshold (See Fig 2).

Based on our analysis, a threshold of 0.175 represents a good tradeoff between precision and recall, resulting in 29.7% of messages being categorized as relevant and an accuracy of 71.2% with respect to the ground truth. But we also need to think about how this threshold will affect the interface and the ability to effectively use the system: it may be more meaningful to minimize either false positives *or* false negatives, and depending on how strict a threshold is applied this will impact the amount of information visible. For instance, a user doing an exhaustive search may be more interested in reducing false negatives (high recall), whereas a user doing a more exploratory search may be more interested in reducing false positives (high precision). In general practice the threshold should be tuned for the context and demands of the situation, but since a threshold of 0.175 results in a relevance set of almost exactly the same size as that found in the ground truth (29%) we use this as the default value in our system.

One might note that our definition of relevance misses messages which are topically related to the speech but which occur after more than a delay of one minute. Indeed, in our development of the ground truth we did find many messages that were germane to a speech topic but were posted at later minutes. We contend that integrating a temporal aspect into our definition of relevance can help an analyst tune into the *immediate* response to an event, but that relaxing the temporal aspect (e.g. having a longer time window) could be useful for other tasks such as exploring the ongoing conversation around a particular topic.
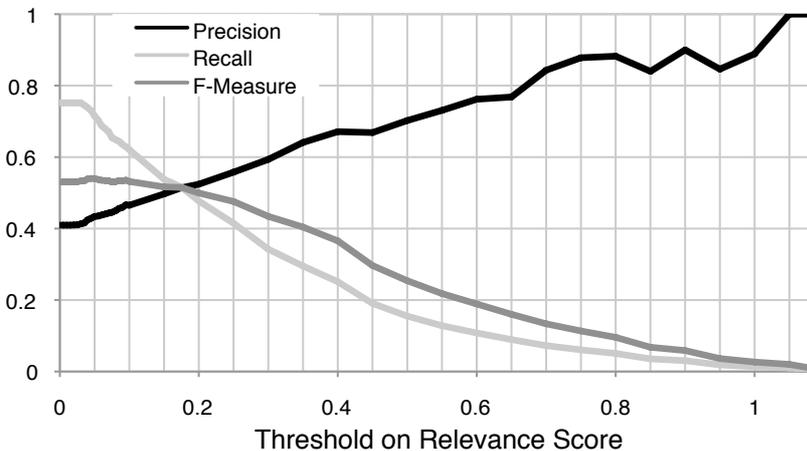

Fig 2 A plot of precision, recall, and F-measure for varying relevance score thresholds. A message with a relevance score below the threshold was classified as not relevant.
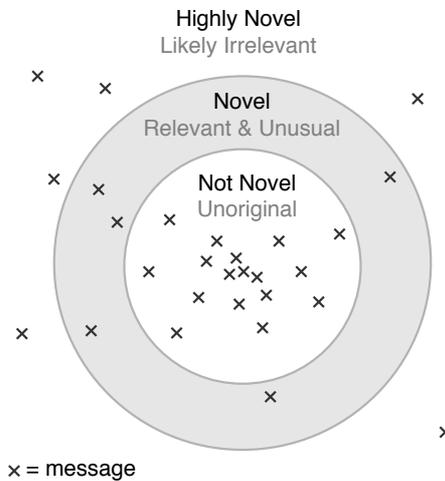
Highly Novel
Likely Irrelevant

Novel
Relevant & Unusual

Not Novel
Unoriginal

× = message

Fig 3 A conceptual diagram showing the relationship between message novelty and utility. Those messages in the middle gray band may be interesting in the sense that they are unusual but still relevant to the discussion3

Novelty

In the context of social media, "unusual" may manifest itself as more unique or *novel* messages when compared to other social media messages provided that the messages are still relevant to the content of the event (see Fig 3). Definitions of "newsworthiness" and "news values" in journalism, our use scenario for the Vox system, often espouse the importance of the *unusual* or *unexpected* in the selection criteria for what becomes "news" [9, 10]. We incorporate this concept into our system by developing a message novelty metric, which can be used to direct attention toward what may be more unusual contributions.

Here we define the novelty of a message in relation to the other messages sent during the same time interval. In this sense novelty can also be thought of as a measure of the *conversational relevance* (i.e. how (dis)similar is a message to all of the other messages being shared at a particular point in time). According to our definition, novelty is then a measure of how different a message is from what everyone else is talking about. We compute novelty 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 aggregation interval of the event (again one minute). 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 novelty using the cosine similarity of term vectors as:

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

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 "interesting" range of novelty for the filter presented in the interface by thresholding novelty values between a minimum and maximum value as suggested in Fig 3.

We evaluated our method for computing message novelty in order to assess whether the novelty score corresponds to human judgments of novelty. We also wanted to characterize what would be effective upper and lower thresholds for determining the "interesting" range of novel messages. This range defines unusual contributions that are not completely irrelevant.

We sampled the State of the Union dataset for one "anchor" message per minute (70 messages total) and 30 "comparison" messages from each of those same minutes (2100 messages total). Both the anchor and comparison messages were selected randomly for each minute. The anchor message in each minute was manually compared to each of the comparison messages and a binary similarity evaluation was recorded for that pair (i.e. the pair of messages was judged as similar or not similar). Thus, for each anchor there were 30 similarity ratings (1 or 0), which in turn were averaged and subtracted from one to compute a novelty score for the anchor message in relation to the other messages sent during that same time interval.

The manual comparison was based on a *semantic* similarity between the messages rather than looking at a lexical similarity. In instances where the anchor message consisted of several clauses, if any of the clauses was similar to a comparison message the pair was evaluated as similar. Messages that contained the same topic (e.g. health care reform) but were different in their meaning were not considered similar. For instance, for the anchor message, "*He's getting defensive. Not backing down on defending the stimulus. Good, but that'll turn away indeps*" a message rated as similar was "*Obama says 2 million jobs created via recovery act. Hmmm.*" because of the reference to the stimulus, and a message rated as not similar was "*I love how animated VP Biden is.*" because there is no semantic similarity to the anchor.

We computed the Pearson correlation coefficient between the automatically computed novelty scores and the manually computed average novelty scores for the anchor messages. We found there to be a correlation ($r = 0.279$, $p < 0.05$) indicating that there is a statistically significant correlation between how a human judge rates novelty based on semantic similarity and how our algorithm rates novelty based on lexical similarity. This indicates that even a relatively simple method based on keyword term vectors produces results that are consistent with how a human rater judges novelty.

To determine practical workable values for upper and lower novelty thresholds we assume a reasonable reduction in the amount of messages in the interface for the novelty filter would be about 10%. We thus chose the upper and lower thresholds to be 0.99 and 0.95, which results in 11.6% of messages being labeled "novel". Similarly to the relevance thresholding, for the current application we felt that this proportion of messages in the filter was appropriate, however, we could have also included facilities in the interface to tweak the proportion of messages filtered based on thresholds allowing an analyst more leeway in self-defining what was the "interesting" range of novelty scores.

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 understanding of the *polarity* (i.e. positive versus negative) of the social media reaction to the event content. 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 applied a supervised learning algorithm (language model) trained with 1900 manually tagged messages randomly sampled from the State of the Union dataset. In this case each message was tagged by a single human coder. The coding schema that we used was based on three categories: *objective* (e.g. factual and free from personal feelings or interpretations), *positive* (e.g. a positive evaluation, opinion, emotion, or speculation), or *negative* (e.g. a negative evaluation, opinion, emotion, or speculation). In order to simplify the schema we elected not to have additional categories for messages that contained both positive and negative sentiment, or for neutral messages – which while making the scheme easier to apply for coders resulted in the loss of some precision in coding these other types of messages. We found the best performance using a language model including all n-grams of length less than or equal to four. The classifier resulted in a 10-fold cross validated accuracy of 63.5%.

In an attempt to improve the accuracy further by using more reliable training data we undertook to have multiple raters per message and to also assess the interrater reliability of the training data. In this case, we randomly selected a different set of 750 messages from the State of the Union dataset. The same coding scheme was used as before. Each of three coders independently applied the sentiment

schema to each message. The majority vote category was taken as the final rating for each message. Coders were fluent in English and lived in the U.S. or Canada for at least 8 years; they were familiar with the cultural context of the dataset. We computed the Fleiss $\varkappa$ for this ground truth as 0.706, indicating a fair level of reliability and agreement among the coders [1]. At the same time, this value of kappa indicates that this is a difficult task for even highly motivated and careful human coders, and that the inherent ambiguity in the training data is a challenge to training an accurate classifier. Similar to above we then developed a language model based on this new training set. We found the best performance using a language model including all n-grams of length less than or equal to four. The classifier resulted in a 10-fold cross validated accuracy of 63.4%, essentially equivalent to the results we found for the larger but less reliable 1900 message training set. In looking at the contingency tables for misclassification errors we found that most of the error is concentrated in negative messages misclassified as positive, and positive messages misclassified as negative. Thus while sufficient for giving an overall impression of sentiment, the classifier still fails on difficult cases such as those 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. These results generally indicate that having more highly reliable training data is not the limiting factor in improving the classification accuracy for sentiment polarity.

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 users exploring the content and response to the event. 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 [15], 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 [21] or words [3] could be implemented and integrated into our data processing pipeline.

To compute the document frequency portion of the 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 [3].

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 4a. 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 4b). Aggregate response analysis is enabled by three views: volume graph (4e), sentiment timeline (4f); and the keywords component (4g). These views are all aligned to the video timeline (4c) and the topic timeline (4d) 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.

Fig 4 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)

Content Component

The content component (Figure 4b) 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 4c) that allows start, pause and nonlinear navigation of the video steam. On the right side of Figure 4b, the interface shows the set of messages about the event that were 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.

Filtering Module

The filtering module, shown in Figure 4a, 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 (4b), as well as the aggregate statistics in the message volume graph (4e) and sentiment timeline (4f). All these components update interactively with the filter. In our current implementation, filtering does not change the keyword pane (4g).

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, (6) messages classified as novel, and (7) messages with positive or negative sentiment. The system also allows for filtering out quotes, retweets or links. In Figure 4, for example, the active filters are the keyword "budget" and "no retweets".

Topic Timeline

We included a topic segmentation timeline (Figure 4d) 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 4c). 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. For the purposes of the evaluation we present later, we opted to produce the topic segmentation manually so as to provide the highest quality experience to the users.

Message Volume Graph

The message volume graph (Figure 4e) 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.

Sentiment Timeline

The sentiment timeline (Figure 4f) shows an aggregate of the sentiment response for each minute of the event, as derived by the sentiment analysis procedure detailed previously. 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 number of positive and negative 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 mes-

sages in the message list (Figure 4b) 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). During the design process we spoke to prospective journalist users who 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.

Keywords Component

The keywords component (Figure 4g) depicts the salient keywords extracted over time. 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 five minutes worth of space, that word's score is the sum of its tf-idf scores for all of those five 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 from 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.

Exploratory Study

We designed and executed an exploratory evaluation of Vox Civitas to assess its effectiveness in a journalistically motivated sensemaking scenario. Journalism is just one use-case for Vox Civitas though we believe there is much to be learned from studying specific contexts and user populations (i.e. 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 addressed the following research questions:

- How useful and effective was the tool for journalists in generating story ideas and reporting on the broadcast 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.

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 U.S. We logged participants' actions with the interface and recorded open-ended survey responses. We identified interactions 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

---

[1] http://sm.rutgers.edu/vox

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 experiment were persistently displayed next to the tool. The instructions to participants specifically read, "Imagine that the State of The Union just occurred are you're using this tool to find stories to pitch to a national news editor. Come up with at least two story angles which, with some more reporting, you think will make good news stories". We intentionally left the task somewhat open-ended in order to allow participants leeway in how they chose to employ the tool. Interactions with the interface (mouse hovers, clicks, time spent using the tool) were logged to give an indication of feature usage.

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.

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".

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.

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 [13]. 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 novel contributions, or "decisive moments" which draw heavy audience response [2, 10]. 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).

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.

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 novelty, 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.

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 goal in this use case is not so much to provide rigorous assessment and decision support about hypotheses (i.e. deductive reasoning), but rather to spur the divergent and creative generation of hypotheses, insights, and questions for follow-up activities (i.e. abductive reasoning) [19].

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 a design that provided for a sensemaking schema, thus organizing the information visually according to cues expected to be of interest (topic, magnitude, sentiment, novelty) to the user.

Vox Civitas obviates the initial phases of the sensemaking process (data collection and schema generation) and allows analysts to focus on 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.

Our evaluation of relevance, novelty, and sentiment revealed several nuances that could help the design of future visual analytics applications. For example, our definition of relevance, though useful for defining a temporal notion of importance, may not be appropriate to all situations. Other visual representations (e.g. streamgraphs) could be incorporated to better depict overlapping and ongoing "relevant" conversations. Our assessment of sentiment classification found that more reliable training data did not increase overall classification accuracy. Future avenues to explore for improving accuracy could involve collecting *more* training data since we only had 750 messages, or in linking to and incorporating dictionaries of slang such as the Urban Dictionary.

Fig 4 depicts the State of the Union event filtered for the keyword "budget" – looking at the volume graph one can see three distinct peaks of activity. The third peak corresponds to what one might expect, people are commenting on the budget because Obama just mentioned it, however the first two peaks are different in nature. Looking carefully at the messages there one can see that oftentimes the *same* message was sent out by different accounts – indeed these different accounts are all linked implicitly in that they are related to the military (i.e. the accounts correspond to Air Force or other military bases). It would then appear that the U.S. military was using Twitter to "spam" the State of the Union by sending out the same message from multiple accounts simultaneously. This vignette draws attention to the need for integrating network analysis into the tool. Ideally such network analysis would help show the relationship of Twitter accounts not only to each other, but also to similar content, and to time.

In our evaluation of both the computational enablers and the interface we focused on a single event, the State of the Union address, in a distinctly journalistic context. There is however a huge range of events, and analytic contexts, where an adapted version of Vox Civitas could still be met with substantial utility. For example, one can imagine a use case of Vox Civitas (or the analytic methods we propose) where the goal is to reason directly about the event content, rather than focus on the response to it.

In addition, different types of events, and maybe even those that do not even have a video channel may also be interesting to examine. For instance, social media related to the FDA (Food and Drug Administration) recalls, alerts, or outbreaks, could be investigated by analysts to help determine how individuals are connected, or how an outbreak is evolving over time. Such investigation does not need to be limited to formal "events", but rather can be applied to longer term evolving issues that could benefit from temporally-driven investigation.

Conclusion

Increasingly, broadcast and other media events can be associated with a set of postings from various social media channels (e.g., Facebook and Twitter) about the event. Posted in real time, the social media content is roughly aligned with the multimedia content of the event, as we have shown above. This new channel of meta-content about the event offers, at the same time, an opportunity to reason about the media content, and about the nature of the response to the content by social media users.

In this context, we presented a tool to support media professionals interested in the social media response to a broadcast event by collecting, analyzing, aggregating and visualizing content from one major broadcast event, the U.S. 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. We intend to pursue an in-depth case-study approach to understanding the tool's efficacy and how it can be iterated upon to be more useful in a wide variety of contexts.

An interesting question for future work would be how to support signal-level analysis of the multimedia content for the event using the social media data, and how to integrate the video and textual content analysis in one analytic framework. The "social media signal" for any broadcast event greatly enhances the availability of textual descriptors for the event content. While it is not yet clear how the temporally aligned social media content relates to the video content for different types of events (sports, speeches, TV dramas), it is clear that it can provide significant cues that are often hard to reliably extract from the video – like, for example, when a goal was scored in a football match.

[1]   Artstein, R. and Poesio, M. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, *34* (4). 555-596.

[2]   Clayman, S. Defining Moments, Presidential Debates, and the Dynamics of Quotability. *Journal of Communication*, *45* (3). 118-146.

[3]   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).

[4]   De Choudhury, M., Sundaram, H., John, A. and Seligmann, D.D., What Makes Conversations Interesting?: Themes, Participants and Consequences of Conversations in Online Social Media. in *Proc. WWW*, (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 *Proc. 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]   Harcup, T. and O'Neill, D. What is News? Galtung and Ruge Revisited. *Journalism Studies*, *2* (2). 261-280.

[11]   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). 9-20.

[12]   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).

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

[14]   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).

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

[16]   Mertens, R., Farzan, R. and Brusilovsky, P., Social Navigation in Web Lectures. in *Proc. Hypertext and Hypermedia*, (2006).

[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]   Pike, W., Stasko, J., Chang, R. and O'Connell, T. The science of interaction. *Information Visualization*, *8* (4). 263-274.

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

[21]   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).

[22]   Sakaki, T., Okazako, M. and Matsuo, Y., Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. in *Proc. WWW*, (2010).

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

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

[25]   Shamma, D.A., Shaw, R., Shafton, P.L. and Liu, Y., Watch What I Watch: Using Community Activity to Understand Content. in *Proc. MIR: Workshop on Multimedia Information Retrieval*, (2007).

[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]   Thomas, J. and Cook, K. (eds.). *Illuminating the Path*. IEEE, 2005.