Matisse: A Visual Analytics System for Exploring Emotion Trends in Social Media Text Streams

Chad A. Steed*, Margaret Drouhard*, Justin Beaver*, Joshua Pyle* and Paul L. Bogen II*
*Computational Sciences and Engineering Division
Oak Ridge National Laboratory, Oak Ridge, TN 37831
Email: csteed@acm.org, mdrouhard@acm.org, beaverjm@ornl.gov, jmpyle771@gmail.com, plbogen@google.com

***Abstract*—Dynamically mining textual information streams to gain real-time situational awareness is especially challenging with social media systems where throughput and velocity properties push the limits of a static analytical approach. In this paper, we describe an interactive visual analytics system, called Matisse, that aids with the discovery and investigation of trends in streaming text. Matisse addresses the challenges inherent to text stream mining through the following technical contributions: (1) robust stream data management, (2) automated sentiment/emotion analytics, (3) interactive coordinated visualizations, and (4) a flexible drill-down interaction scheme that accesses multiple levels of detail. In addition to positive/negative sentiment prediction, Matisse provides fine-grained emotion classification based on Valence, Arousal, and Dominance dimensions and a novel machine learning process. Information from the sentiment/emotion analytics are fused with raw data and summary information to feed temporal, geospatial, term frequency, and scatterplot visualizations using a multi-scale, coordinated interaction model. After describing these techniques, we conclude with a practical case study focused on analyzing the Twitter sample stream during the week of the 2013 Boston Marathon bombings. The case study demonstrates the effectiveness of Matisse at providing guided situational awareness of significant trends in social media streams by orchestrating computational power and human cognition.**

I. Introduction

Social media systems, such as Twitter, give users open forums to broadcast their thoughts and experiences continually to a global audience. Due to widespread utilization of mobile technologies, these systems continue to transform public discourse and set trends and agendas in most facets of society. The rising value of these systems for global situational awareness is apparent as we see journalists increasingly turn to social media and similar online streams to detect and investigate breaking news events and identify sources of expertise. In these situations, the ability to detect changes in sentiment, emotion, opinion, and behavior in social media streams for the well-being and safety of the public is both feasible and desirable.

Despite the potential advantages, social media stream analysis is a challenge because of the large and rapidly changing content delivered in the form of semi-structured textual information—one of the fastest growing data types in the big data era. Understanding this information requires the ability to grasp key trends and interactively drill-down to increasingly detailed views of the data. More specifically, an interactive visual analytics system is needed that orchestrates the strengths of humans and computational machinery. Computational power is utilized to efficiently process the streaming content. On the other hand, human intuition, creativity, high-bandwidth visual processing, and background knowledge are engaged via interactive visual interfaces.

In light of these challenges, we have developed an interactive visual analytics system, called Matisse (see Figure 1), to explore trends and associations in social media streams. The current work significantly expands upon Matisse's introduction [1] with more material related to key sentiment/emotion analytics and an expanded case study. Supported by a flexible stream data management framework and guided by automated sentiment/emotion analytics, Matisse allows data-driven, human-directed analysis of social media streams via interactive visualizations. Matisse encourages a multi-scale analytical workflow that begins with high-level overviews and progresses to increasingly detailed visualizations, which may include accessing the raw data records.

The remainder of this paper provides a description of the following key components in Matisse: (1) text stream data management; (2) automated analytics for positive/negative sentiment estimation and more detailed emotion classification; and (3) multiple coordinated visualizations that support multi-scale drill-down exploration via human interaction. Then, we present a practical case study using tweets captured from the Twitter sample stream during the Boston Marathon bombings.

II. Related Work

A. Visual Text Analytics for Social Media

Due to the popularity of social media services, several visual analytics systems for exploring textual microblog content are present in the literature. For example, Wanner et al. present a survey of significant research contributions and insights for streaming text visual analytics, focusing on the development of techniques for event detection [2]. Suntinger et al. examined a particular event detection method for the Event Tunnel [3]. The Event Tunnel's clustering methods allow users to identify patterns, including causal relationships, in historical data.

Some visual analytics systems focus on event or entity detection and topic modeling. Krstajic et al. described stacked

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan)

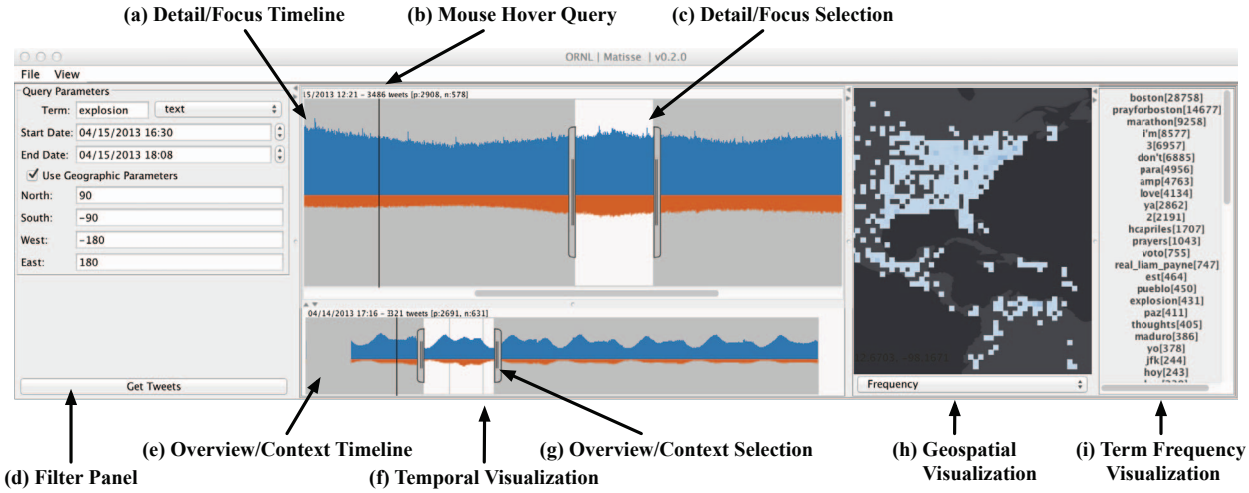


Fig. 1. Matisse's main interface features a highly interactive visual canvas for exploring text stream information in temporal, geospatial, and term-frequency visualizations. The visualizations are linked using a coordinated model and a filter panel allows detailed record retrieval. In this Figure, tweets captured during the week of the Boston Marathon Bombing are visualized using bivariate timeline graphs (blue bars indicate positive tweet frequency and orange bars indicate negative tweet frequency). The selected time range reveals a noticeable spike in positive/negative sentiment during the event.

time series and radial tree graphs to explore the relationships between entities in large-scale news streams using entity recognition and correlation [4]. Dörk et al. introduced the Visual Backchannel concept for exploring online conversations about large scale events in Twitter [5]. The Visual Backchannel provides multiple coordinated views in a web-based system with an image cloud, a stacked time series graph, and a unique spiral representation of activity between people. More recently, Dou et al. [6] created the LeadLine system, which focuses on automatic identification of events using topic modeling, event detection, and named entity recognition techniques. The LeadLine system allows users to interactively explore the events identified. Although useful for topic exploration and capable of representing multiple time series, neither LeadLine nor Visual Backchannel provide stream analysis or drill-down analysis as Matisse does.

Other systems draw inspiration from the ThemeRiver [7] technique for visualizing topic evolution in text collections. The TextFlow tool, introduced by Cui et al., combines topic mining methods and text visualizations to assist users in progressively analyzing topic evolution [8]. EventRiver is designed to explore temporal trends in text streams such as weblogs (blogs) and news corpora [9]. EventRiver clusters documents to find events of interest and visualizes the events as a "river" that encodes patterns and relationships between events. Although the temporal visualizations in Matisse resemble those of EventRiver, Matisse focuses on bivariate encodings, multiple linked visualizations, and it targets the format and nuances of textual social media content. The EvoRiver system introduces several interactive visualization techniques inspired by ThemeRiver for the exploration of topic cooperation and competition over time [10]. The OpinionFlow system uses an opinion diffusion model to approximate opinion propagation among Twitter users as well as an opinion flow visualization to convey diffusion [11]. The FluxFlow system implements a data processing system similar to Matisse, but with a different focus—to reveal and analyze anomalous information spreading on social media (e.g., rumors and misinformation) [12].

B. *Sentiment Analysis*

Sentiment analysis techniques have evolved to categorize even small snippets of text in fine-grained detail. Alm et al. implemented a modified Winnow update rule-based classifier using the SNoW learning architecture to classify sentences into fine-grained emotional categories [13]. The semantic features, weighting of WordNet emotion words, and the supervised classifier yield strong results. The techniques introduced by Aman and Szpakowicz [14] and Strapparava and Mihalcea [15] rely on a combination of an emotion lexicon and corpus-based features to classify text snippets into six emotion categories. The former implemented a Support Vector Machine classifier, while the latter used a Naïve Bayes classifier. The "Attitude Analysis Model" (@AM) classifies emotions using nine categories [16] and relies on the assumption that the meaning of a sentence is determined by the meanings of its word or phrase components. In addition to emotion classification, dimensional models, including the Valence-Arousal-Dominance (VAD) model used within Matisse, have performed well for sentiment analysis tasks. Kim et al. used unsupervised learning techniques to evaluate the respective merits of two models for emotion—a categorical model and a dimensional model [17]. Of the methods evaluated, categorical non-negative matrix factorization and dimensional methods performed best.

Other methods isolate sentiment toward specific entities or topics. Engonopoulos et al. introduced a technique to classify the sentiment of each word in a document using Conditional Random Fields [18]. The sequential nature of the method facilitates the identification of sentiment toward specific entities. Most recently, Van de Kauter et al. demonstrated a method to detect both explicit and implicit sentiment in text from financial news articles and identify the topic or subject of the sentiment [19]. Matisse's emotion classifier does not explicitly identify the topic of the sentiment, but the coordinated views allow the analyst to interactively discover and investigate these relationships.

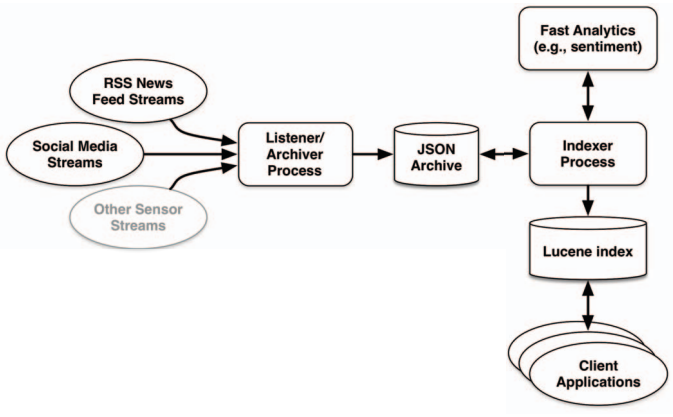


Fig. 2. Matisse incorporates a robust data management system that consists of a listener process and an indexer process. Stream information is transformed into a Lucene index for processing by client applications. During indexing, analytical algorithms are applied to the textual information.

III. Text Stream Processing Framework

Efficient management of streaming textual information is vital to any interactive analysis and visualization system. Matisse is supported by a customized stream management framework (see Figure 2) that includes a robust listener process and a stable open source search engine. The listener collects entries and periodically serializes the data at a customizable time interval (e.g., every second, minute, hour) into an archive of JavaScript Object Notation (JSON) formatted files. In a separate process, an indexer service continuously monitors the JSON archive. When new information is detected, the indexer extracts relevant fields and adds it to a Lucene index. After indexing, the data is available to the various analysis components through programming interfaces. Although the stream processing system is capable of consuming any textual information (e.g., RSS news feeds, blog streams), the current work focuses on trend analysis for the Twitter sample stream. Twitter's sample stream provides a 1% random sample of all public tweets as well as filtered stream queries over specific geographic areas and/or keywords of interest.

To feed multi-scale displays, the system automatically computes summary information for the streams. These summaries are based on temporal, geospatial, and textual context. For example, temporal summaries are created using statistical metrics (e.g., frequency, averages, variance) calculated for tweets collected in evenly distributed time interval objects, or time bins (e.g., every minute, hour). These time bin summaries are used in the linked visualizations to graphically represent trends in the information flowing through the system. The summary bins can also be calculated on-the-fly for use in our interactive multi-scale visualizations. Similar data structures are calculated and stored to describe the geospatial and term-frequency information in each time bin.

IV. Automated Sentiment and Emotion Analytics

The Matisse system leverages automated sentiment and emotion analytics to reduce the complexity of the raw information through classification. In the following subsections, we describe two analytical capabilities: (1) Positive/negative sentiment prediction, and (2) fine-grained dimensional emotion classification.

A. Positive/Negative Sentiment Prediction

To estimate positive or negative sentiment, we use an approach similar to Go et al. [20]. The process for moving from raw text to a feature vector begins with converting all characters to lower case. Then, all tokens beginning with the '@' character are removed as these tokens represent to whom a tweet is directed and generally carry no sentimental information. The next step is to identify any instance of three or more of the same character in a row and reduce the segment to just two of the same character. For example, the following tweet, 'Iiii looooveeee yooooouuuu' becomes 'Ii loovee yoouu'. Character repetitions of three or more are replaced to prevent tokens like 'little' from becoming 'litle'. The process decreases to two characters to preserve the ability to distinguish the emphatic 'yooooouuuu' from the non-emphatic 'you'. Next, common contractions are identified (can't, won't, etc.) and replaced with their standard version (cannot, will not). The next step removes bracketing characters ('(', ')', '[', and ']') and compresses extra whitespace. The process then removes all URLs and replaces them with the token 'URL'. Finally, any remaining punctuation is removed.

The processed string is then segmented into tokens, stemmed using Porter's English stemmer, and filtered for stopwords using a custom stopword list that leaves in common sentimental words that are often included in standard stopword lists such as 'want', 'not', 'should', and 'could'. From these token lists, a vector is generated consisting of the tokens in the filtered input. The process does not use counts, as the 140 character limit of a tweet makes token reoccurrences rare. Notable differences between our approach and Go et al. [20] include the discarding of '@' tokens, expansion of contractions, use of a full stemmer, a customized stopword list, use of Boolean features as opposed to numeric features, and no reliance on bigrams.

To train our classifier, we utilized the publicly available training set mentioned by Go et al. [20]. We discovered that while their work produced positive/neutral/negative sentiments, they do not provide training data for neutral tweets. For the current work, we opted to produce positive/negative sentiments only. The processed tweets from the Go et al. data set are used to train both a Python and Java Naïve Bayes classifier and a Java Maximum Entropy Classifier. For Python, we utilized *nltk* [21] along with *scikit-learn*. For Java, we used *MALLET* [22] and *MinorThird* [23]. In our tests, the Python-based Naïve Bayes classifier achieved a 90% accuracy rate. Under Java, a Naïve Bayes classifier performed at a 79% accuracy rate and a maximum entropy classifier performed at an 82% accuracy rate. These results are comparable to the 83% classification accuracy reported in Go et al. [20] for unigram and bigram feature sets under a Maximum Entropy classifier.

B. Emotion Classification using Machine Learning

To provide a more detailed mood estimate, we developed a new method to classify tweets from a broader range

Lucene is an open source search engine http://lucene.apache.org.

Twitter sample stream information is available at http://dev.twitter.com.

Information about *scikit-learn* is available at http://scikit-learn.org

of emotions by merging the presence of tokens with their significance from both a textual frequency and emotional perspective. Generalizing the emotion associated with a tweet is challenging due to the limited amount of textual data, the tendency to include abbreviations and symbols to represent broader concepts, and the lack of ground truth data to form the basis for a model [24]. Furthermore, defining a practical and repeatable method for quantifying the emotion associated with a tweet presents difficulties that are related to the size and nature of the tokens. We have addressed these challenges with an evaluation of various candidate methods of machine learning, classification schema, emotion representation, term/phrase significance, feature selection, and ground truth acquisition.

The Affective Norms for English Words (ANEW) model [25] serves as the basis for our emotion quantification. ANEW is a dictionary of terms that have been quantified (interval scale) using three dimensions: (1) Valence describes positivity/negativity, (2) Arousal indicates the excitability in the text, and (3) Dominance depicts the level of assertion. The presence of any of the 1,034 ANEW terms within a tweet forms the basis for its quantified emotional context. The overall emotional score for a tweet is estimated by averaging the Valence, Arousal, and Dominance (VAD) scores for all ANEW words in the text. To simplify the representation, the VAD scores are discretized over their score range. Valence is represented by the *unpleasant/neutral/pleasant* categories, Arousal is represented by the *subdued/neutral/active* categories, and Dominance is represented by the *dominated/in-control* categories. Thus, our approach to the VAD estimation of emotion for a tweet results in one of 18 possible combinations of the VAD discretized categories For example, a VAD categorization of *pleasant/neutral/dominated* indicates that the average raw VAD scores for the ANEW terms in the text map to the *pleasant* Valence category, the *neutral* Arousal category, and the *dominated* Dominance category.

Using the Term Frequency-Inverse Corpus Frequency (TF-ICF) term weighting method [26], the most significant terms are determined in each textual record. TF-ICF calculates term frequency in a document by transforming the text into a vector space model [27], and approximating each term's importance based on weights determined from independent large text corpora. The outcome of applying TF-ICF is a degree of significance associated with each of the tweet's tokens, which can be used to determine the importance of any ANEW terms with respect to the overall document. The most significant terms and ANEW emotion scores are merged as a feature set that is representative of the content and emotion associated with each textual record.

We transform the tweet text into a collection of name-value pairs called a *feature set*. These feature sets are combined with the emotion category labels to produce *examples* that guide the learning algorithms in the creation of a classification model. First, the text is vetted for the presence of ANEW terms, and the VAD scores are averaged and discretized to obtain an overall categorical representation of emotion. Next, the text is converted into a vector space model and TF-ICF is used to identify the most significant terms. The VAD categories and most significant terms are combined in a feature set that estimates the emotion associated with the text.

Our machine learning approach estimates emotion using

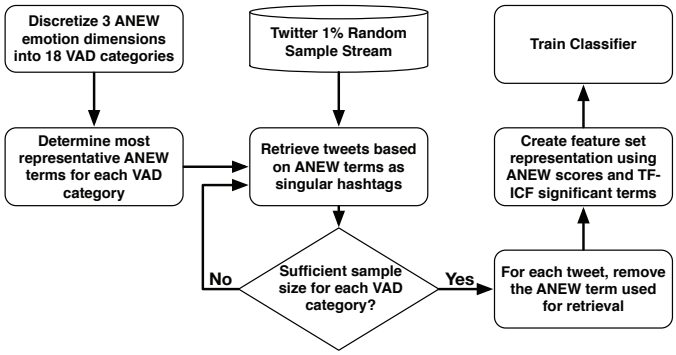


Fig. 3. Notional view of the emotion classifier training process.

the 18 discretized VAD categories and its novelty lies in the learner training method. The challenge in training a classifier is selecting features that accurately represent the emotion associated with the underlying text, and retrieving a set of textual examples that are unique to a specific emotion class. The training process leverages features from both statistical and emotional text analysis, and it samples data based on the most pure examples of the various emotion classes. As shown in Figure 3, we acquire labeled examples of emotion classes by selectively retrieving tweets with a specific emotion class explicitly encoded as a hashtag. For example, to obtain labeled examples of the VAD categories *pleasant/subdued/dominated*, we first find the set of terms from the ANEW dictionary that map to these discretized VAD ranges. Using this subset of ANEW terms, a set of tweets are retrieved wherein at least one of these ANEW terms is represented within the tweet as a singular hashtag. The reasoning behind this approach is that the hashtag emphasizes the word, which in turn emphasizes the emotion associated with the word, providing the least uncertainty in the automated labeling process. This process is repeated for the 12 most representative ANEW terms in each VAD category (representativeness is determined by the euclidean distance to the centroid of the valence-arousal-dominance space for each discretized category) until a sufficient data representation is achieved. We take the additional step of removing instances of the ANEW keyword used for retrieval from each tweet to prevent bias towards the retrieval criteria. Once the set of emotion-labeled tweets have been retrieved, feature sets are derived for each using the ANEW quantification of VAD emotions and the inclusion of significant terms as determined through TF-ICF, and then a classifier is built to predict the emotion in new unlabeled tweets. Based on prior natural language processing experience, we selected a maximum entropy learner from the *MinorThird* machine learning library [23] for multi-class emotion classification.

We have evaluated the performance of the classification model using tweets captured from the Twitter sample stream during the Spring of 2013. Tweets with singular hash tags of the most representative ANEW keywords of the 18 VAD categories were acquired until a sufficient sample size was reached for each category. A minimum sample size of approximately 100 tweets was expected for each VAD category, with most categories comprised of greater than 350 examples. We performed a 10-fold cross-validation, where the training data is randomly subsampled into 10 groups. A process was repeated 10 times where a classifier was built from 9 of the 10 random

subsets and then applied to the remaining subset, which was treated as the validation data set. The misclassification error rates were then averaged to get an overall sense of the error rates associated with each class. Table I lists the results of this analysis. Each of the VAD classes is listed with the average number of labeled examples used in training each of the 10 cross-validation models, and the resultant average error rate. There appears to be some association between quantity of exemplars and error rate as the VAD classes trained with fewer than 350 examples exhibit higher error rates. However, for classes with higher example counts, there appears to be no correlation between number of examples and error rate, indicating that the VAD and textual features associated with each class may be the driver for variations in the generalization performance. Overall, the classifier performs well in the cross-validation with an average error rate across all classifications of 0.0329 and a balanced error rate across all classes of 0.0344. We interpret these results as an initial validation that our approach to training a high-fidelity emotion classifier is sound, but we recognize that further study is needed for this claim to be conclusive.

TABLE I. EMOTION CLASSIFIER CROSS VALIDATION ERROR RATES

Valence	Arousal	Dominance	No. Examples	Error Rate
Pleasant	Active	Dominated	739	0.0279
Pleasant	Active	In Control	703	0.0335
Pleasant	Neutral	Dominated	551	0.0194
Pleasant	Neutral	In Control	952	0.0498
Pleasant	Subdued	Dominated	484	0.0295
Pleasant	Subdued	In Control	804	0.0378
Neutral	Active	Dominated	731	0.0221
Neutral	Active	In Control	528	0.0334
Neutral	Neutral	Dominated	381	0.0258
Neutral	Neutral	In Control	397	0.0203
Neutral	Subdued	Dominated	268	0.0430
Neutral	Subdued	In Control	444	0.0259
Unpleasant	Active	Dominated	739	0.0306
Unpleasant	Active	In Control	556	0.0291
Unpleasant	Neutral	Dominated	621	0.0319
Unpleasant	Neutral	In Control	278	0.0469
Unpleasant	Subdued	Dominated	193	0.0814
Unpleasant	Subdued	In Control	367	0.0315

V. INTERACTIVE MULTI-SCALE VISUALIZATION TECHNIQUES

Matisse leverages interactive data visualizations to allow human-centered visual exploration of trends in social media streams. The visualizations leverage the information produced by the previously mentioned analytical methods to render temporal, geospatial, term frequency, sentiment, and emotion visualizations. These separate views are coordinated such that interactions in one display are systematically propagated to the other displays. Furthermore, the system supports drill-down investigations from high level overviews to detailed record listings using multi-scale representations. The geospatial view (see (h) in Figure 1) presents a heat map for the selected time range and supports basic zoom and pan operations. The color scale used in the map represents grid cells with higher tweet counts as darker and more saturated shades of blue and lower tweet counts as lighter and less saturated shades of blue. The term frequency view (see (i) Figure 1) shows the top ranked terms for the selected time range. Selecting individual terms populates the filter panel text field for retrieval of detailed tweet information. In the remainder of this section, we introduce the key details of the timeline and emotion classification visualization techniques. For more detailed information about Matisse's visualization capabilities, the reader is directed to our previous work [1].

A. Overview+Detail Temporal Visualization

Matisse's aggregates the summary statistics for a user-defined unit of time (seconds, minutes, hours, etc.) to form a time series. As shown in Figure 1, this time series information is graphically encoded in a visualization that represents the summary metric as a bivariate bar chart. The bar chart shows the frequency of positive sentiment tweets as blue bars (top) and the frequency of negative sentiment tweets as orange bars (bottom) with a common baseline in the center representing zero. The blue and orange bars can be added together to obtain the overall frequency for the time intervals. As the user moves the mouse cursor (see (b) in Figure 1), detailed summary information is shown above the timeline as a hover query.

In Figure 1, Matisse displays two timeline views: (1) an overview/context view and (2) a detail/focus view. The bottom view (see (e) in Figure 1) is the overview timeline, which provides an overall summary of the selected variable(s) for the entire time series. In this view, the time bin summaries are condensed to fit the width of the window panel. Therefore, the time duration represented by each bar is determined by the width of the display and the bin dimensions are recalculated each time the window is resized. The user can select a time range in the overview (see (g) in Figure 1) by using the mouse to drag a rectangle. When the drag operation is complete, the scrollable details timeline (see (a) in Figure 1) is regenerated with the most granular time bin information for the context selection time range. In Figure 1, the detail timeline shows minute level statistics. The user can also make time range selections in the details view (see (c) in Figure 1), which will propagate to the other linked visualizations.

B. Emotion Classification Visualization

As shown in Figure 6, Matisse offers a supplemental view for visually investigating the emotional classification of selected text items. This visualization relies on the VAD emotion classifier described in Section IV-B and the 18 VAD categories. Each point in the scatterplot represents a quantity of tweets assigned to one category for a particular time period. The default scatterplot view uses a standardized point radius and has average Dominance values along the x-axis and average Arousal values along the y-axis. Every tweet in a time bin from the main Matisse timeline view is assigned a category by the emotion classifier, and each of the tweets assigned to any given category will be aggregated and represented as a summary point in the scatterplot. The emotion scatterplot supports zoom and pan interactions. When the user hovers the mouse cursor over points within the scatterplot, additional information about that point is displayed. Furthermore, the point and other points belonging to the same category are highlighted with a user-defined color.

The user may adjust either of the axes to represent the ranges of values for Valence, Arousal, and Dominance. The user may also change the display to map the radius of each point to the total number of tweets assigned to the emotion category of a given tweet timeline bin (see Figures 6(a) and

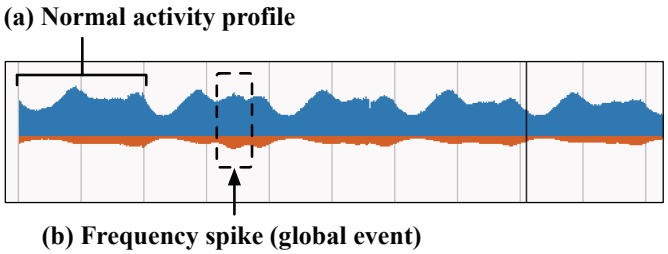


Fig. 4. After initially loading the Twitter sample stream tweets into the overview timeline, we notice a frequency spike (b) on the second day by comparing it to the profile of a typical time series (a). The spike deviates from the normal flat afternoon pattern and indicates a significant global event.

6(b)). In addition, the user may adjust both color and opacity for each category as well as the highlight color. By default, point color is determined by color families mapped from Valence of the category. Each color family has two members: a lighter and less saturated color, representing an "InControl" Dominance category, and a darker, more saturated color that represents a category with the "Dominated" field. Finally, the user may interactively filter the categories displayed (*pleasant/neutral/unpleasant*, *active/neutral/subdued*, *dominated/incontrol*) and save a numerical summary of the VAD classification of tweets represented in the scatterplot.

VI. Case Study: Analysis of the Twitter Sample Stream During the Boston Marathon Bombing

To illustrate the effectiveness of the Matisse visual analytics system, in this section we describe a practical case study involving a real-world scenario. The scenario focuses on the discovery of significant trends in tweets captured from the Twitter 1% sample stream during the week of the Boston Marathon bombings (April 14–20, 2013). The objective is to illustrate how Matisse's interactive interface supports intuitive drill-down investigations that progress from clues in high-level overviews to dynamic queries of increasingly detailed evidence using a flexible, human-centered methodology. We present the case study from the perspective of an analyst tasked with monitoring the Twitter sample stream in real-time to identify and investigate significant trends.

First, the processed Twitter stream index is visualized (see Figure 4) to show the positive and negative tweet frequencies for the entire time series in a condensed bivariate graph. The first day exhibits a normal timeline pattern (see Figure 4(a)) for the Twitter sample stream. That is, activity increases from a minimum in the early morning hours to a peak at noon. Then, activity drops slightly to a steady state in the afternoon and rises again in the evening until midnight. Based on our experience, deviations from this typical pattern, such as the spike highlighted in Figure 4(b), indicate the occurrence of globally significant events. In this case, the deviation occurs during the afternoon on April 15, 2013 with a peak of 4,235 tweets (3,207 positive and 1,029 negative) at approximately 5:15 p.m. We note that the spike is visible in both the positive and negative tweet time series.

Boston Marathon bombings details are available at http://en.wikipedia.org/wiki/Boston_Marathon_bombings

Eastern Time Zone (UTC-04:00) references are used in this case study.

With evidence of a significant event, we drill-down to gain more insight. In Figure 5, we select a time range in the overview timeline roughly centered on the day of the observed spike. Due to the linked view model, this selection causes the details timeline to render the minute-level time bin information for the selected range, which confirms the increase in activity during the afternoon with higher granularity. Using the mouse hover query, we observe a positive to negative tweet ratio of approximately 3:1. Visual inspection of the detailed time series indicates that this ratio is generally consistent in the selected time range. The geospatial view and the term frequency views reveal additional clues about the event prompting the activity spike. In the geospatial view, darker blue bins in the northeastern U.S. suggest the epicenter. Furthermore, top terms related to Boston refine our location estimation and terms such as "marathon" and "prayers" suggest possible ties to an event and feelings of compassion, respectively.

As shown in Figure 1, we drag an additional time range of interest in the detail view to restrict the information rendered in the geospatial and term frequency views to a smaller time span (approximately 2 hours centered on 5 p.m.). The refreshed geospatial view reveals a high concentration of tweets around the Boston, MA region. In addition, the refreshed term frequency view provides textual clues. For instance, terms with "boston" and "marathon" suggest that something occurred at the Boston Marathon and terms such as "explosion" and "prayer" suggest a disastrous scenario.

In order to grasp the emotional reactions of people tweeting about the trend, we open the VAD emotion panel for a period surrounding the frequency spike. Figure 6 shows a comparison of the emotion analysis for tweets during the same afternoon period on April 15 and April 20, 2013. The time period analyzed on both days is 4:01 p.m. to 6:31 p.m. On April 20, the day after the bombing suspect, Dzhokhar Tsarnaev, was apprehended, the overwhelming majority of tweets are classified as pleasant or neutral valence. In contrast, the analysis of April 15 shows a higher-than-average number of tweets classified as unpleasant. The unusually high ratio of negative tweets supports our theory that a serious emergency has occurred on April 15. Using the category filters to investigate further, we observe that there are noticeably more tweets categorized as subdued arousal on April 15 than on April 20. A summary of the most divergent categories, both broad (valence and arousal) and specific, between the two days is shown in Table II.

TABLE II. Most significant VAD categories for afternoons of April 15 and April 20, 2013.

Category	4/15/13 4:01 - 6:31 p.m.	4/20/15 4:01 - 6:31 p.m.
Total tweets:	592092	414565
Arousal Subdued (%):	57.65	61.80
Valence Unpleasant (%):	16.04	11.38
Pleasant-Subdued (%):	26.88	30.07
Pleasant-Subdued-Dominated (%):	10.52	14.04

With more details, our initial indication of atypical Twitter activity evolves into a rough hypothesis that people are discussing an event in the Boston area that may involve explosions and may be related to the annual Boston marathon. To test our hypothesis, we continue to drill-down to more detailed information by double-clicking on the term "explosion", which populates the query term text field in the left filter panel as shown in Figure 1. Then, we click the "Get Tweets" button

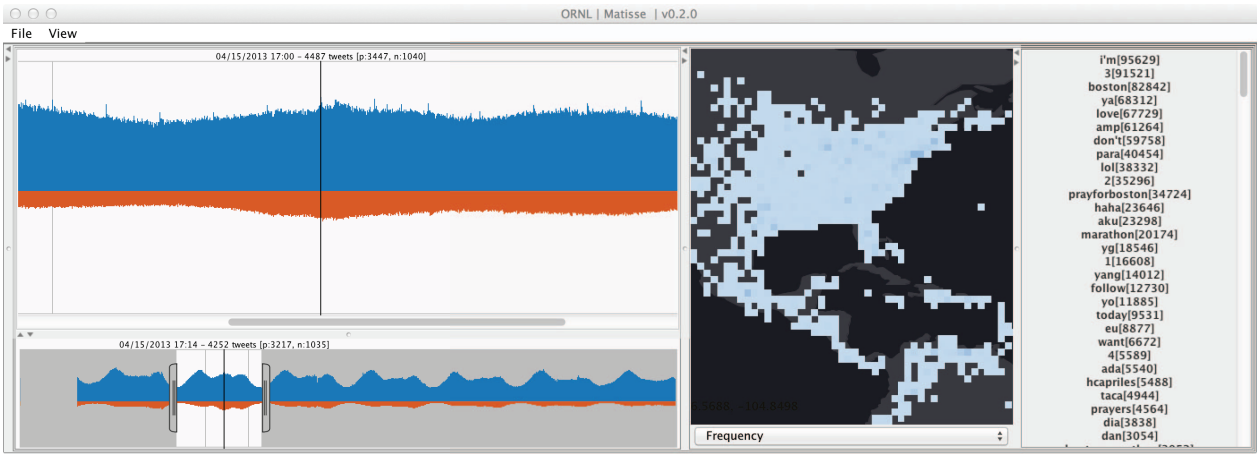
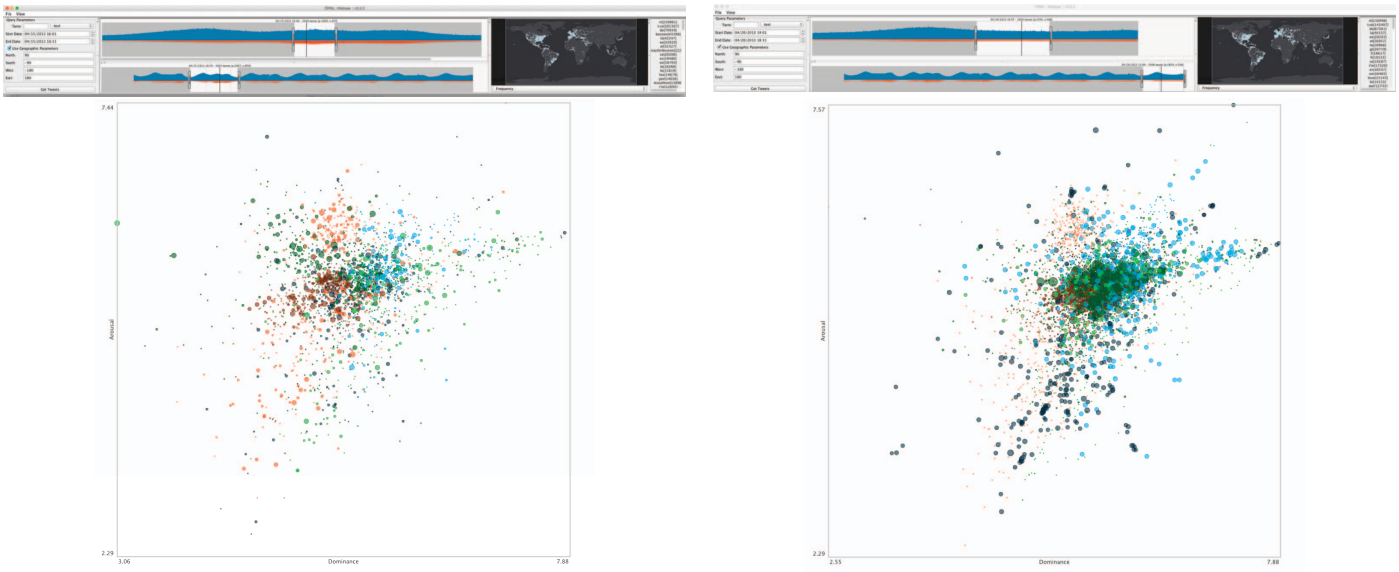


Fig. 5. A frequency spike during the afternoon of April 15 prompts the selection of that day in the bottom overview timeline. The detail timeline is rendered using the most detailed summary information (minute intervals), providing more granularity. The spike in both positive and negative sentiment is still visible and the geospatial and term-frequency views indicate possible linkages to the Boston, MA area.



(a) VAD Emotion View for April 15, 2013

(b) VAD Emotion View for April 20, 2013

Fig. 6. The VAD emotion views for April 15 and 20, respectively, using a radius proportional to the number of tweets per tweet bin. The same time period (4:01 p.m. - 6:31 p.m.) is shown for both days. Tweets from April 15 show a noticeably higher ratio of "unpleasant"-labeled tweets (orange points) and significantly fewer "pleasant" (blue points) sentiments than those from April 20.

to list all tweets containing this term during the selected time period. The resulting listing of tweets (not shown due to Twitter data usage restrictions) includes several descriptive responses to the Boston Marathon bombing and confirms our hypothesis about the details of the event that has caused the increase in Twitter activity.

Other terms in the view suggest the presence of additional global trends. The term "jfk" appears because a fire at the JFK Presidential Library around the same time as the bombing sparked fears of a related explosion. The fire was determined to be unrelated to the explosion, but the public's rapid reaction illustrates the societal impact of social media.

The exploration of the timeline, geospatial, and VAD emotion views, as described above, also point to another significant global trend. The heavy concentration of tweets in Caracas and other cities in Venezuela indicate a significant event may be occurring in that area. The keywords "hcapriles," "maduro," and "voto" ("vote") suggest that the event is related to the April 14 special election in Venezuela following the death of Hugo Chávez. Nicolás Maduro and Henrique Capriles were the two primary candidates, and allegations of electoral fraud spurred protests in the capital and other locations throughout the country. Top keywords such as "cacerolazo" ("protest"), "derrotar" ("to defeat"), and "lucha" ("fight"), which can be explored through the "Get Tweets" feature, align with our initial assessment of a significant event in Venezuela. The keywords and listing of tweets indicate that civil society groups organized protests surrounding the election. Despite multiple events, it is straightforward with a visual exploration system like Matisse to formulate and confirm hypotheses from high-level overviews to detailed data investigations.

This case study illustrates the effectiveness of Matisse at delivering data-driven, human-guided analysis that starts with high-level overviews and systematically descends to precise raw information through multiple levels of detail. Built upon a coordinated multiple view interaction model, the system fosters creative and flexible exploration of social media text streams in a manner that leverages both computational and human strengths.

VII. Conclusion

Effective situational awareness of large and rapidly changing social media text streams requires a balanced approach between abstraction and elaboration. Although detailed visualizations provide high fidelity, they are impractical for understanding global trends and associations. On the other hand, aggregated visualizations can highlight broad trends but are incomplete without access to the detailed records in context. The Matisse visual analytics system harmonizes these opposing requirements using coordinated visualizations that allow drill-down investigations of trending topics through multi-scale views. Future work includes new visualization techniques that tie sentiment/emotion to topics and using interactive human interaction to label emotions for active learning. In addition, we are planning more comprehensive evaluations of the emotion classification algorithm, which is widely regarded as an extremely difficult task given the lack of ground truth data [24]. Given the ever-increasing volume and velocity of social media and other related text streams, the intelligent orchestration of both human and computational strengths is a necessity rather than an option. Matisse is a practical exemplar of the type of visual analytics system that is required to enable trend discovery and guided investigations for these important channels of societal issues and global events.

Acknowledgment

This work is based upon work supported by Oak Ridge National Laboratory LDRD project No. 6427.

References

[1] C. A. Steed, J. Beaver, P. L. Bogen II, M. Drouhard, and J. Pyle, "Text stream trend analysis using multiscale visual analytics with applications to social media systems," in *IUI Workshop on Visual Text Analytics*, Mar. 2015, pp. 1–8.

[2] F. Wanner, A. Stoffel, D. Jäckle, B. Kwon, A. Weiler, and D. Keim, "State-of-the-art report of visual analysis for event detection in text data streams," in *EuroVis-STARs*, 2014, pp. 125–139.

[3] M. Suntinger, H. Obweger, J. Schiefer, and E. Groller, "The event tunnel: Interactive visualization of complex event streams for business process pattern analysis," in *Proceedings of the IEEE Pacific Visualization Symposium*, March 2008, pp. 111–118.

[4] M. Krstajic, F. Mansmann, A. Stoffel, M. Atkinson, and D. Keim, "Processing online news streams for large-scale semantic analysis," in *Proceedings of the IEEE International Conference on Data Engineering Workshops (ICDEW)*, March 2010, pp. 215–220.

[5] M. Dörk, D. Gruen, C. Williamson, and S. Carpendale, "A visual backchannel for large-scale events," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1129–1138, 2010.

[6] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou, "LeadLine: Interactive visual analysis of text data through event identification and exploration," in *Conference on Visual Analytics Science and Technology (VAST)*, 2012, pp. 93–102.

[7] S. Havre, B. Hetzler, and L. Nowell, "ThemeRiver: visualizing theme changes over time," in *IEEE Symposium on Information Visualization*, 2000, pp. 115–123.

[8] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong, "TextFlow: Towards better understanding of evolving topics in text," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2412–2421, 2011.

[9] D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. Keim, "EventRiver: Visually exploring text collections with temporal references," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 1, pp. 93–105, 2012.

[10] G. Sun, Y. Wu, S. Liu, T.-Q. Peng, J. Zhu, and R. Liang, "EvoRiver: Visual analysis of topic coopetition on social media," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1753–1762, 2014.

[11] Y. Wu, S. Liu, K. Yan, M. Liu, and F. Wu, "OpinionFlow: Visual analysis of opinion diffusion on social media," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1763–1772, 2014.

[12] J. Zhao, N. Cao, Z. Wen, Y. Song, Y.-R. Lin, and C. Collins, "FluxFlow: Visual analysis of anomalous information spreading on social media," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1773–1782, 2014.

[13] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: Machine learning for text-based emotion prediction," in *Proceedings of the ACL Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 579–586.

[14] S. Aman and S. Szpakowicz, "Using roget's thesaurus for fine-grained emotion recognition," in *IJCNLP*, 2008, pp. 312–318.

[15] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proceedings of the ACM Symposium on Applied Computing*, New York, NY, USA, 2008, pp. 1556–1560.

[16] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Recognition of affect, judgment, and appreciation in text," in *Proceedings of the ACL Conference on Computational Linguistics*, 2010, pp. 806–814.

[17] S. M. Kim, A. Valitutti, and R. A. Calvo, "Evaluation of unsupervised emotion models to textual affect recognition," in *Proceedings of the NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 2010, pp. 62–70.

[18] N. Engonopoulos, A. Lazaridou, G. Paliouras, and K. Chandrinos, "ELS: A word-level method for entity-level sentiment analysis," in *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, 2011, pp. 12:1–12:9.

[19] M. Van de Kauter, D. Breesch, and V. Hoste, "Fine-grained analysis of explicit and implicit sentiment in financial news articles," *Expert Systems with Applications*, vol. 42, no. 11, pp. 4999–5010, 2015.

[20] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, pp. 1–12, 2009.

[21] S. Bird, E. Loper, and E. Klein, *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[22] A. K. McCallum, "MALLET: A machine learning for language toolkit," 2002. [Online]. Available: http://mallet.cs.umass.edu

[23] W. W. Cohen, "Minorthird: Methods for identifying names and ontological relations in text using heuristics for inducing regularities from data," 2004. [Online]. Available: http://minorthird.sourceforge.net

[24] R. Zafarani and H. Liu, "Evaluation without ground truth in social media research," *Communications of the ACM*, vol. 58, no. 6, pp. 54–60, 2015.

[25] M. M. Bradley and P. J. Lang, "Affective norms for english words (ANEW): Instruction manual and affective ratings," University of Florida, Tech. Rep. Technical Report C-1, 1999.

[26] J. W. Reed, J. Yu, T. E. Potok, B. A. Klump, M. T. Elmore, and A. R. Hurson, "TF-ICF: A new term weighting scheme for clustering dynamic data streams," in *Proceedings of the 5th International Conference on Machine Learning and Applications*, 2006, pp. 258–263.

[27] S. K. M. Wong, W. Ziarko, and V. V. Raghavan, "On modeling of information retrieval concepts in vector spaces," *ACM Transactions on Database Systems*, vol. 12, no. 2, pp. 299–321, 1987.