Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media

EmotionWatch: Visualizing Fine-Grained Emotions in Event-Related Tweets

Renato Kempter and **Valentina Sintsova** and **Claudiu Musat** and **Pearl Pu**
School of Computer and Communication Sciences
Swiss Federal Institute of Technology (EPFL), CH-1015, Lausanne, Switzerland
{renato.kempter, valentina.sintsova, claudiu-cristian.musat, pearl.pu}@epfl.ch

Abstract

Spectators are increasingly using social platforms to express their opinions and share their emotions during big public events. Those reactions reveal the subjective perception of the event and extend its understanding. This has motivated us to develop a system to explore and visualize volume, patterns, and trends of user sentiments as they evolve over time. Previous work in sentiment analysis and opinion mining has addressed these issues. But the majority of them distinguish only two polarity categories, leaving a more detailed and insightful analysis to be desired. In this paper, we suggest using a fine-grained, multi-category emotion model to classify and visualize users' emotional reactions in public events. We describe EmotionWatch, a tool that constructs visual summaries of public emotions, and apply it to the 2012 Olympics as a test case. We report findings from a user study evaluating the usability of the tool and validating the emotion model. Results show that users prefer a more detailed inspection of public emotions over the simplified analysis. Despite its complexity, users were able to effectively grasp, understand, and interpret the emotional reactions using EmotionWatch. The same user study also pointed out few design improvements for the future development of analogous systems.

Introduction

While television allows people to watch big public events, such as the Olympic Games, movie awards or political debates, social media lets spectators, participants and other event followers with various cultural backgrounds interact and engage with each other. Along with some facts and description, they share their emotional reactions about an event. These intertwined emotions of the public, if summarized and reconstructed, reflect the subjective perception of the event, and can open up new perspectives for all stakeholders involved. Spectators could compare their own emotions with the feelings of others to better understand an event. Social scientists could construct and validate hypotheses about the emotions and their causes, while journalists could find valuable moments and reactions about an event. And marketers could detect patterns and trends of social opinions to improve marketing campaigns.

This inspiration motivated us to build a tool to help users gain a quick and succinct overview of how the public react to a big social event, and to track how those reactions evolve over time. While text-based summarization tools could achieve this goal (Sharifi, Hutton, and Kalita 2010), to further enhance this process for users, recent tools, e.g. TwitInfo (Marcus et al. 2011), employ information visualization techniques to reveal patterns and trends of sentiments in visual forms. However, the majority of these tools classify sentiments expressed in tweets using two levels of granularity, positive and negative (Marcus et al. 2011; Zhao, Wickramasuriya, and Vasudevan 2011; Diakopoulos, Naaman, and Kivran-Swaine 2010). We believe that the polarity classification does not provide enough distinctions for our feelings. It cannot distinguish whether people are deeply involved with or happy about the event, or whether they are worried about or compassionate for the event participant. Moreover, research has shown that visual representations of sentiment at polarity level alone can mislead and irritate users (Marcus et al. 2011): a negative event, such as an earthquake, can provoke positive tweets with messages aimed at supporting the affected community. Literature on emotion research shows that human emotions are far more subtle, fine-grained, and expressive than a model of simple polar opposites (Ekman 1992; Scherer 2005; Parrott 2001; Plutchik 2001). Using polarity for the representation of emotion can thus only be a starting point. Tools that can more *accurately and comprehensively* classify and visualize *emotional reactions* have yet to be developed.

We present EmotionWatch – a tool that automatically *recognizes* emotions in social media and gives a *visual summary* of the public reactions such events cause. After carefully examining four well-grounded emotion models from psychological research, we settled on the Geneva Emotion Wheel, version 2.0 (GEW) (Scherer 2005), a tool for evaluation of emotional reactions at much higher levels of expressiveness. While other models of basic or primary emotions contain up to 8 emotion classes (either Ekman's (1992), Plutchik's (2001) or Parrot's (2001)), GEW covers 20 discrete emotion categories most frequently stated in self-assessments. This greater number enables us to more accurately analyze subtle details. Also, unlike the other models mentioned, this set of emotions has an equal number of positive and negative emotions. Only with this model we can distinguish such different

positive reactions as enjoyment, love and laughter caused by, for example, political speeches.

In this paper, we describe the detailed design and validation process for EmotionWatch, using the 2012 Olympics as a test case, and make the following contributions:

1. Our system maps the emotional content of tweets into 20 GEW-based emotion classes, whose meanings users can readily understand and distinguish. To the best of our knowledge, EmotionWatch is the first application for exploring and interpreting users' feelings at events with a fine-grained multi-category emotion representation.
2. Our system lets users explore exceptional moments, shown by *peaks* representing intense emotional reactions, *aggregates* of multi-colored sparks representing controversial reactions, and *outliers* representing emotional anomalies.
3. Our system employs a novel visualization technique combining a traditional time line approach, with a real-time and dynamic animation to vivify public emotions. It can support both visualization of collective emotional responses to an event and the comparison of responses to different actors involved. It not only summarizes the emotions expressed in social media, but also attempts to evoke viewers' emotional responses.

To validate the contributions, we conducted a formative evaluation. Results show our users could successfully understand, perceive, and interpret the expressed emotions. The emotions shown by EmotionWatch were found to be consistent with the video of an event. Users could recognize and distinguish emotions, as well as discern special moments such as peaks and aggregates. Most importantly, they reported a strong preference for multi-category emotion model over the simplified polarity representation. The study also revealed several insights on how to further improve the system, which we will discuss in detail.

The rest of the paper is structured as follows. We begin with an overview of related work. We continue with a description of the data we would like to visualize for our test case. Next, we give details about our visualization technique and the interfaces proposed for data investigation. We proceed with an evaluation of our system, discuss future work and offer a conclusion.

Related Work

This section reviews related visualization techniques and systems for summarizing public events using microblogging platforms.

Data Visualization Our data consists of the values for multiple emotion categories at different points of time, i.e. we visualize multivariate time-series data. Miksch and Schumann (2011) reviewed multiple visualizations for this type of data. They distinguish static visualizations, where the data for all time moments are shown simultaneously. One example would be a 3D Kiviat Tube (Tominski, Abello, and Schumann 2005), where one axis represents time and each perpendicular 2D cut shows the diagram for a time point data; another example would be 2D stacked graphs, such as the ones used in Muse system (Hangal, Lam, and Heer 2011) to show the timeline of the sentiments. Such static representations offer an easier overview and comparison of the data in different time points. However, visualizing many categories in this way can complicate the ability to read and distinguish emotions in the current time interval. In the context of showing sentiments and emotions, researchers used simplified static visualizations, e.g. a one-axis timeline using colors to show the most prominent emotion of each moment in time (Liu, Selker, and Lieberman 2003; Chen et al. 2008). We believe that using this approach in isolation is not helpful for aggregating emotions from multiple documents, because a variety of emotions can be present in a given moment. Thus, we blend such simplified static visualization with the dynamic one detailing all categories of active emotions, where the dynamic part reflects a cumulative summary of emotions in the current time interval.

The summary of affective content can be presented in different ways. For example, Gregory et al. (2006) suggested a rose plot with values for each emotional dimension shown on a separate petal. Alternatively, when emotions are viewed as points in dimensional space, the emotion summary can be formed by mapping those dimensional values into specific visualization parameters, such as color and shape of a bubble in AffectAura (McDuff et al. 2012); or by placing the values of all the tweets into the emotional plane.[1] The idea of individual document visualization can be also used with the categorical emotion model. For example, WeFeelFine shows each document as a particle with the color corresponding to the stated emotion (Kamvar and Harris 2011). There were not yet studies on how users perceive such emotion summary visualizations. Thus, we suggest another way to represent the cumulative summary – a modified star plot (also known as a radar chart) (Chambers 1983), which can provide a compact and clear structure of an emotion distribution.

Summarizing Events Using Social Media Using tweets related to an event, a textual summary can be extracted automatically with a result similar to a news report (Nichols, Mahmud, and Drews 2012). For recurring events, such as the Olympic Games or the Oscar Movie Awards, existing knowledge about the internal workings of the event can be employed. For example, sub-events, such as touchdowns in American football or goals in soccer, can be detected from a message stream and classified with given labels (Chakrabarti and Punera 2011). Systems visualizing a tweet stream have proposed different statistical cues to help identify important moments during events (together with the corresponding information) (Shamma, Kennedy, and Churchill 2010). Such systems reconstruct and visualize the event structure from the tweets and allow content analysis and an overview of the flow of the event. Yet, they do not answer exclicitly the question of how people are reacting to or interacting with the event.

Several summarizing systems have included the visualization of sentiments to help investigate personal reactions towards time-framed events. TwitInfo (Marcus et al. 2011) was developed to detect peaks in tweet streams, automat-

[1] www.csc.ncsu.edu/faculty/healey/tweet_viz/

ically add labels to those peaks, and visualize them in a timeline-based display for browsing and exploration. Additionally, it displays information on opposite polarity sentiments in the form of a pie chart. Yet, such a chart has been found unreliable because of differences uncovered between the expected sentiments of an event and the summary of tweet sentiments. We see the presentation of emotions instead of sentiments, and distinction of multiple emotion classes, as a way to avoid such a problem.

Personal reactions and opinions given during television programs can hint what viewers find more interesting and engaging to watch. SportSense (Zhao, Wickramasuriya, and Vasudevan 2011) reveals television watchers' sentiments on major sports events in real time. It detects the percentage of positive tweets minus the negative tweets and presents this in a simple plot in time. Such a representation compares sentiment between competing teams, and inspired our idea for comparing opinions about different event participants.

Event summaries from social media can provide benefits for journalistic inquiries. Diakopoulos, Naaman, and Kivran-Swaine designed the Vox Civitas specifically for this goal (2010). Along with content cues, it shows a color-coded sentiment timeline for different aggregate reactions (positive, negative, or controversial). It turned out journalists were especially looking for controversial topics to generate new ideas. We believe that multi-category emotion visualization can add even more details to these controversies.

The Emoto project (www.emoto2012.org), similar to us, focuses solely on a visual representation of the sentiments of tweets, designed for the case of 2012 Olympic Games. However, while Emoto applies only the polarity dimension to categorize tweets, we develop a fine-grained, multi-category approach. Nevertheless, the project's dynamic interface inspired us, especially by the tweets emerging from the display to extend the sentiment summary. We added additional static and contextual cues to understand the overall flow and causes of emotional reactions.

Overall, only few summarization systems put the analysis and summary of people's emotional reactions at the center of their work; rather they presented sentiments as additional dimensions of content. Moreover, all systems showed sentiment in terms of polarity classifications only. Our work addresses the challenge of designing a fine-grained visual representation of emotions, as well as evaluating ease of use and the utility of such an approach.

Application Data for the Test Case

In this section, we describe the input data for the visualization interface, which we will use in our test case of the 2012 Olympic Games. We discuss the processes of tweet collection, annotation with emotions and structuring with respect to the Olympic entities, such as events and athletes. While the suggested methods are specific for the chosen case of the Olympic Games, similar techniques can be designed for other public events by adapting the choice of keywords, specifying the event participants and event time schedule, and tailoring the emotion lexicon to the chosen domain.

Data Collection

During the 2012 Olympic Games, we collected a large dataset of Olympic-related tweets between July 26th and August 14th. The list of search keywords contained references to the Olympic Games (e.g. *Olympics* or *London2012*). All keywords could appear in the tweet text with or without a hashtag (#). This process captured around 36 million tweets. In order to detect English tweets, we applied a language identification tool[2] to all tweets. We excluded hashtags (words starting with #), links and usernames (marked with a handler @) in this filtering process. The final dataset contained 33.2 million English tweets about the Olympic Games.

Emotion Annotation

Having collected the tweets, we aimed to detect the emotions present in each of them using the chosen fine-grained emotion model of 20 GEW categories. Among the attempts to carry out multi-category emotion recognition from tweets, most tried to categorize emotions into several basic categories (up to 8) (Wang et al. 2012; Kim, Bak, and Oh 2012; Mohammad 2012). We chose the lexicon approach of Sintsova, Musat, and Pu (2013), which presents a desirable, fine-grained emotion classification for the field of tweets at sporting events. The lexical approach is commonly acceptable, with the use of emotion lexicons providing the associations of words or phrases with the different emotion categories (Mohammad 2012; Mohammad and Turney 2013; Strapparava and Valitutti 2004). More developed techniques on the basis of such lexicons can incorporate rules that take into account the phrase, sentence and overall textual relationships, as well as modifiers and negations (Neviarouskaya, Prendinger, and Ishizuka 2007). Another alternative could be machine-learning techniques, which have been shown to be successful in the presence of large-scale annotated, even if by hashtags, data (Wang et al. 2012; Mohammad 2012; De Choudhury, Gamon, and Counts 2012). However, their performance in fine-grained emotion model was not yet studied.

Our emotion recognition module employs the $OlympLex$ emotion lexicon, which was created using crowdsourcing techniques (Sintsova, Musat, and Pu 2013). It contains 3193 terms, from unigrams to 5-grams, and uses the desired GEW emotion categories. Each term is assigned to the specific emotion distribution, represented as a normalized 21-tuple with the corresponding values of each of the 20 emotion categories plus a *No emotion* category.

To compute the emotion profile of a tweet, we summed the associated values of each emotion category for all the lexicon terms found in the text. The resulting sums (20 values) capture the emotion profile of the tweet. If a tweet contained no terms from the lexicon, then it was assigned a neutral emotion (*No emotion*). We omited the terms occurred in the tweet text sub-contained in another lexicon term (e.g. if *love you* appears in the tweet, *love* is not counted). We refrained from normalizing the emotion profile for a tweet

[2] code.google.com/p/language-detection/

(e.g. we did not force the sum of all values to be 1), because we wanted to retain information about highly emotional tweets – those containing more emotional terms.

As a result, we discovered that 59.3% of collected English tweets contain at least one emotional term. This confirms our assumption on presence of emotions in Olympic tweets.

Data Structuring

EmotionWatch aims to visually represent the emotional reactions for concrete scheduled events, such as Olympic competitions shown on television. We focused solely on 8 specific events so as to manually guide the process of identifying event- and athlete-related tweets.

Event-Related Tweets We define an event at the Olympic Games as a single scheduled competition, e.g. the final of the balance beam competition in women's gymnastics. The schedule of events during the 2012 Olympics was provided as open data in (The Guardian 2012).

We adapted the commonly used hashtag-based approach. For each of our selected events, using the event name and the name of the discipline, we constructed a list of event-related hashtags. For example, for the balance beam competition we used *#balancebeam*, *#balance*, *#beam*, *#gymnastics*, etc. Event-related tweets contained at least one hashtag from the list and were posted during the event's time-frame. Furthermore, we include all tweets mentioning athletes participating in the event posted during its same time-frame. We took athletes' names from open data (The Guardian 2012).

Athlete-Related Tweets Athletes were one of the main triggers of emotions in Olympic tweets: people cheered for their favorites, shared their impressions on the athletes' performances and worried for the results. Thus, the extraction of these references not only helped us to increase event-related tweet extraction, but allowed us to separate people's reactions towards specific athletes.

People referenced athletes on Twitter in different ways: using a surname, given name, short name or sometimes the full name. These could also be in different orders, and with or without spaces and hashtags. Moreover, people could also link to an athlete's Twitter account using a handler @. We used all these patterns to find athlete-related tweets using the athletes' known full names. Corresponding short names were taken from the Wiktionary.[3] Athletes' Twitter accounts were reconstructed as the first-ranked accounts returned in searches of full athlete names on Twitter. We also excluded any ambiguous references which were entries in the WordNet dictionary[4] or were repeated for several athletes. To test if such an approach is reasonable, we checked 100 randomly chosen tweets marked as athlete-related, and found that 86 of them contained certainly correct references. Overall, 15% of all collected tweets contained references to athletes, while 58% of these athlete-related tweets were detected as emotional by the emotion recognition algorithm.

[3]en.wiktionary.org/wiki/Appendix:English_given_names

[4]wordnet.princeton.edu

Interface Overview

Emotion Wheel

The main visualization tool used in our interfaces is the emotion wheel (Figure 1). It shows the cumulative emotional profile of a given set of tweets. The emotion wheel is a modified form of the radar chart (also named star plot) (Chambers 1983), which is used to display multivariate data with an arbitrary number of variables. It visually represents the values for the 20 categories of emotion in the GEW model (element A, Fig. 1) and the number of tweets in the set (element B, Fig. 1). In the GEW model, emotion categories are presented in a wheel structure, therefore it is sensible to represent our emotional findings using the same, well-studied circular layout. We chose a radar chart as the basis because of its ability to visualize multiple variables in a way that allows visual comparisons between different sets of tweets. The 20 axes, each representing an emotion category, surround the wheel at equiangular distances from each other. The scaled value of each category is plotted as a point on the corresponding axis. A line connecting all 20 points forms the shape in the middle (element C, Fig. 1), and represents the emotional profile found in the current set of tweets. The scaling process includes the division of all 20 emotion values by the maximum value among them. This way, the dominant emotion has a value equal to 1 and is shown with the highest spike running out to the edge of the wheel (Pride on Fig. 1). The values of all the other emotional categories are proportional to it and thus generate smaller spikes. This scaling approach visually highlights the dominant emotion on the wheel, and helps users perceive the other emotions present too.

Compared to the standard radar chart, the emotion wheel visually represents an additional variable that aims to compensate for the value scaling process – the number of tweets in the set. This variable is represented by the outer gray ring

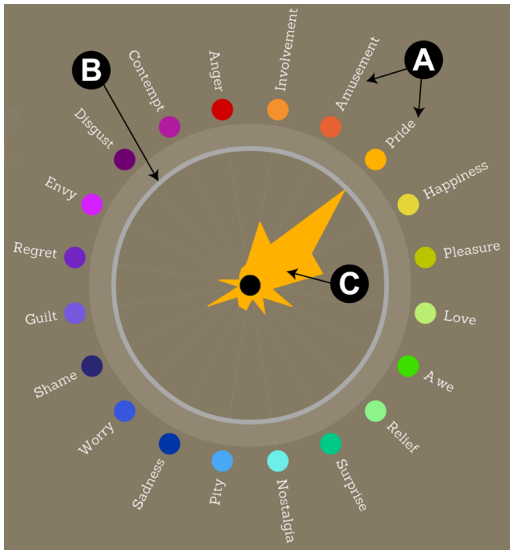

Figure 1: The emotion wheel. A - GEW emotion categories; B - Number of tweets visualized as the ring width; C - Emotion shape visualizing the emotional profile as a star plot.

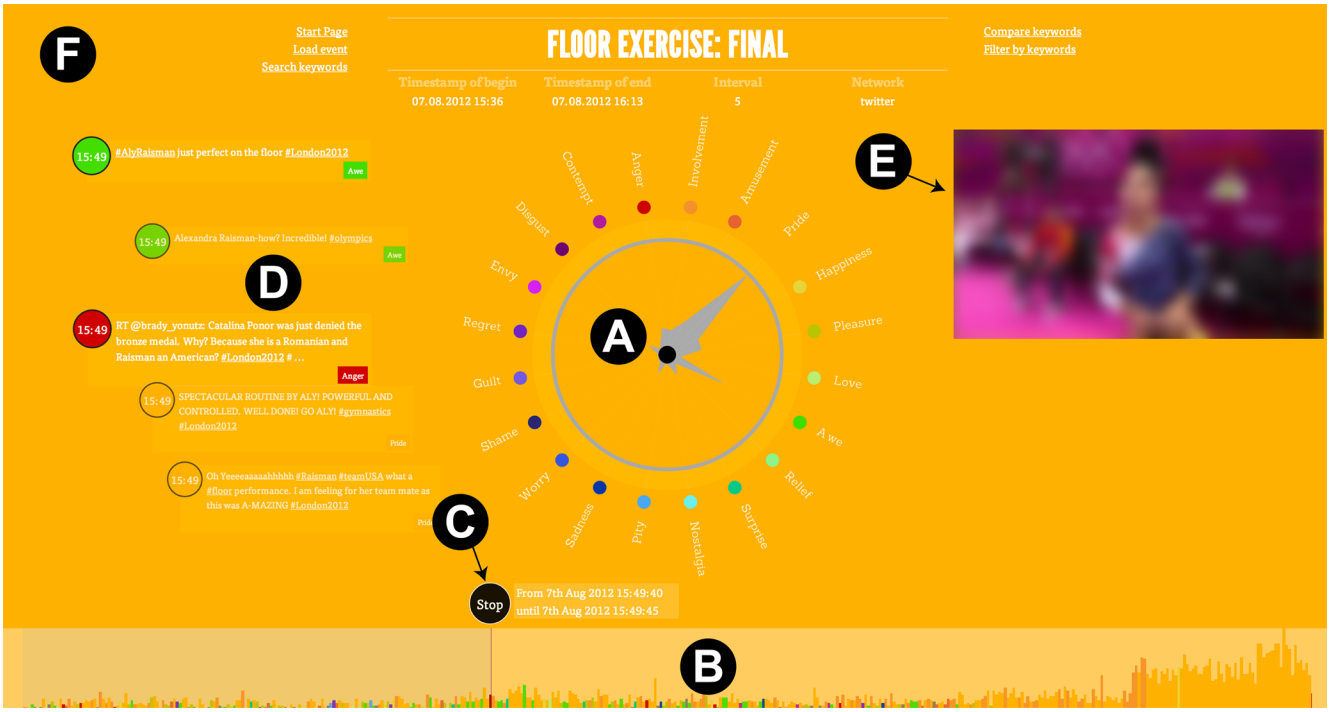

Figure 2: The detailed view showing the Women's Gymnastics Floor Exercise Final. A - Emotion wheel showing the emotion profile of the current time interval; B - Timeline visualizing the emotion flow; C - Button to stop/resume the animation; D - Tweets of the current time interval; E - Video; F - Background with color of the dominant emotion.

(element B, Fig. 1) and its width is proportional to the number of tweets in the set represented. The widest gray ring corresponds to the maximum number of tweets found in all the sets represented. As a further change to the radar chart, we placed a small black circle in the center of the wheel to cover the low magnitude emotion values which could be residual noise of the emotion recognition approach used.

The emotion category names (element A, Fig. 1) surround the wheel arranged towards corresponding axis. Although the original GEW (version 2.0) presented each emotion category using two closely related terms forming an emotional family, e.g. *Sadness/Despair* or *Love/Tenderness*, we use only one term per family on the interface to facilitate user's understanding. Each emotion name appears next to a colored circle representing it. The middle shape (element C) bears the color of the dominant emotion to simplify its recognition. Another possibility is to color the background, instead of the middle shape, to get even more attention to the dominant emotion. We assigned colors to emotions to help distinguish them visually and induce more affective response to the interface. Research on the association between colors and emotions has only been performed using a few of each, without focusing on the separability of the colors for the diverse set of emotions (Kaya and Epps 2004; Simmons 2011). However, several colors are linked to basic emotions, and are widely enough accepted within Western culture to be commonly exploited in applications visualizing emotions (Chen et al. 2008). For example, people usually associate *Anger* with red, *Sadness* with blue, and *Happiness* with yellow. We tried to preserve this allocation of colors for our multiple emotions. We followed the GEW structure and assigned appropriate colors in sequence following the spectrum color wheel. Instead of equiangular color shifts, we readjusted the colors to make them more distinguishable and to reinforce the stated assignment of basic emotions. In order to better dissociate adjacent colors on the wheel, we alternated darker and brighter colors.

Interfaces

Our system consists of two visualization interfaces designed for two different purposes. We present them both below.

Detailed View The detailed view (Figure 2) aims to show the overall emotional reactions during a time-framed event. It is an animated chronological visualization of the emotional responses found in tweets about the event.

The event is split into small, equal time intervals. The tweets in each time interval are grouped together and their emotion profiles in terms of GEW categories are aggregated into an emotion profile for the interval. The visualization dynamically expresses the emotions of each time interval in sequential order for 5 seconds, with an animated linear transition between intervals creating an effect of continuous flow.

The detailed view aims to answer the following questions: What is the *dominant emotion* at a particular point in time? What other emotions are present and to what extent? What is the *context* of these emotions and how are they expressed in tweets? How are emotional reactions *evolving* over time?

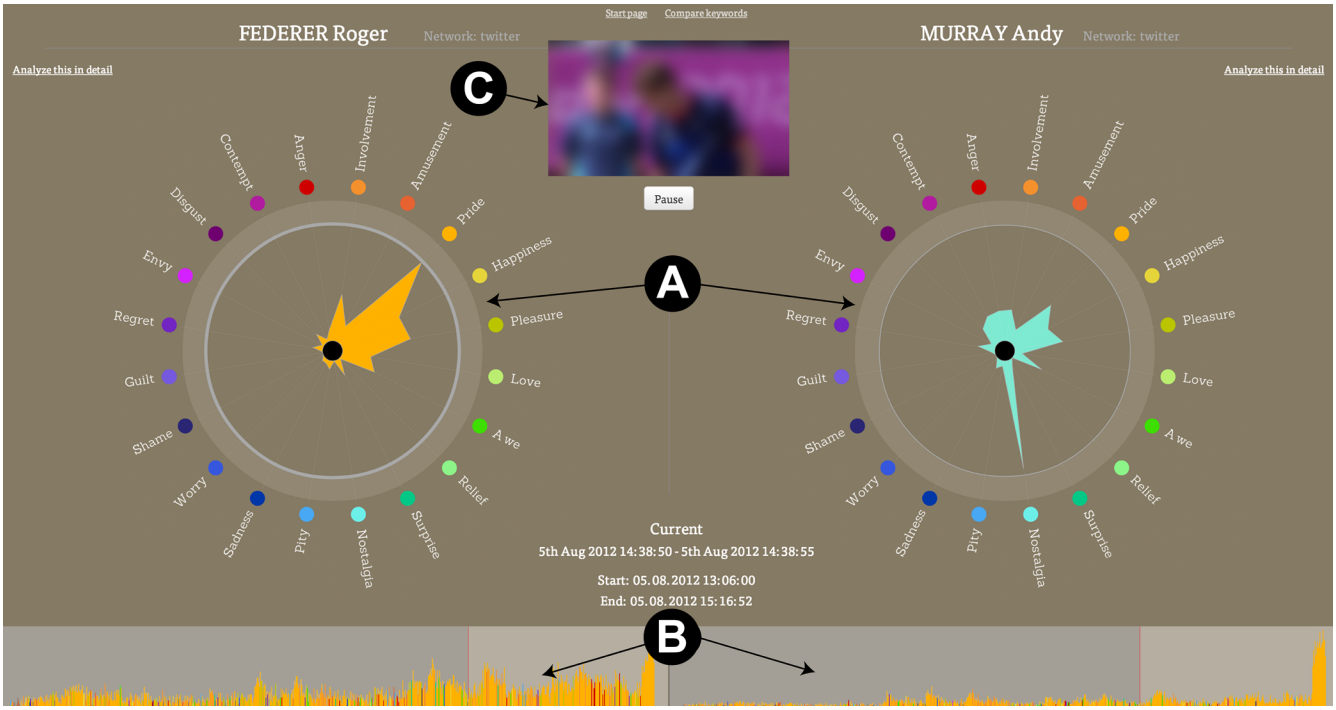

Figure 3: The comparison view showing the tennis final between Roger Federer and Andy Murray. A - Emotion wheels visualizing the corresponding emotion profiles of two athletes; B - Timelines showing the two emotion flows; C - Video.

The emotion wheel in the center of the interface visually represents the emotions in the active time interval (element A, Fig. 2). A timeline at the bottom of the screen (element B) shows all the intervals in chronological order, using bars with heights proportional to the number of tweets in that interval. The highest bar corresponds to the largest number of tweets occurring in any of the intervals. Each bar's color indicates the dominant emotion of its respective time interval. Hovering the cursor over a specific bar in the timeline displays a preview of its corresponding emotion shape on the emotion wheel. It helps users quickly find the moments with the most interesting emotional shapes (e.g. those that have many emotions present).

The timeline serves as a navigation tool: clicking on a bar makes the system jump to visualize the corresponding time interval. The bar representing the displayed interval is highlighted. A box just above the active bar gives additional information about the interval's start and end times. The button attached to this box (element C) can stop or resume the animation, giving the user time to absorb more details on the contextual information. The contextual information corresponds to the active interval and represents a random sample of tweets (area D, Fig. 2) containing emotions. Tweets are animated such that users have the impression of flying through them, simulating a "time travel". Showing actual tweets has two purposes: 1) users can discover what actually happened in the event that caused the emotions present (context), and 2) users can overview how these emotions were expressed. To dig deeper into an event and focus on particular participants, users can filter their names. If available, we show a video recording of the event's television broadcast to provide additional context (element E, Fig. 2). It is synchronized with the visual representation of emotions.

In this view, the page's background color (element F, Fig. 2) changes according to the dominant emotion (instead of the emotion shape's color changing in the wheel). We added this feature to boost attention to the dominant emotion.

Comparison View During an event, emotional reactions can develop around a specific event actor, e.g. an athlete in sports event or a speaker in political debates. The comparison view (Figure 3) allows comparing such emotional reactions using two emotion wheels, each visually representing the emotions of a different set of actors' names or keywords. For example, one could compare the reactions in tweets mentioning athletes from different countries, or investigate how the tweet reactions towards a specific actor contribute towards the overall emotional flow of the event.

The comparison view thus tries to answer the following questions: What are the emotional reactions towards different actors at the same moment in time? How do they differ from each other over the entire event? How do the reactions towards one actor contribute to the overall tone of the event?

EmotionWatch opens the comparison view with a form with which users can select the different keywords or athletes' names they wish to compare during a specific event. The actual visualization interface is split into two halves, each representing one of the two tweet streams for comparison, and similar to, but smaller than, the single detail view. The visualization is animated, showing the emotion wheels

(element A , Fig. 3) of the current time interval at the top and the representative position in the tweet timeline (element B, Fig. 3) at the bottom. The only difference is the absence of tweet texts due to limited screen space. Preview and timeline navigation work in the same way as in the detailed view; except that both wheels always show the same moment in time. Hovering the cursor above a bar in one of the timelines shows the respective emotional shapes on both wheels for that time interval, as does clicking on a bar. For contextual information, a video (element C) is shown if available.

Interface Evolution

The current interface is the result of an iterative design approach. Different prototypes have been created, tested, rejected and selected.

Originally, the emotion wheel included not only the visual representation of the distribution of emotions at a moment in time, but also the information that was later encoded into the timeline (number of tweets in all subsequent time intervals). We considered two ways of presenting the timeline on the watch-like emotion wheel. The first idea was to show the timeline as an additional outer ring, with gradients indicating hot spots (high tweet frequencies) and cold spots (low tweet frequencies). However, because of the colors used for indicating tweet frequency, a color-emotion association would not have been possible. The second idea was to present the timeline as a bar chart (histogram) mapped on the outer sphere of the wheel. User testing revealed that such a tool was too complicated: too much information was packed into one element. Based on the user feedback, we separated continuous event information (the timeline) from the data characterizing specific intervals.

Another variation concerned the shape of the emotion wheel itself. In an early version, the value shown for each emotion was computed relative to the overall maximum emotion value, and not, as in the last version, relative to the dominant emotion of the active time interval. As a consequence, in the original version, the size of the emotional shape strongly depended on the number of the tweets in its respective interval (more tweets resulted in a bigger emotional shape). The pilot user study revealed that this made the visualization less interesting, because the shape was often small and not interpretable, especially at the beginning of an event when only few tweets were arriving. Moreover, a comparison of two emotion wheels would be difficult if one had much smaller emotional shape than the other. With the current version, where tweet frequency is separate from the emotional content, users are able to compare and interpret the two shapes in the wheel and weight them according to the outer ring which indicates the tweet frequency.

User Interface Evaluation

We now demonstrate the proceedings of our user study in which we evaluated the effectiveness of our interface at visualizing the emotional content of the tweets. We performed an in-depth, qualitative, formative evaluation to examine how users interacted with our system, how they interpreted and reasoned with the emotional reactions shown, and which elements they found more important and helpful. Furthermore, we investigated how people conceived our fine-grained model and used color-emotion spectrum.

Method

We recruited 8 graduate students for the user study: 5 of whom watched at least one 2012 Olympic Games event; 5 of whom had a Twitter account; 3 of whom were female. All recruited users were interested in sport in general and had a favorite Olympic discipline to watch. They were all inquisitive about the emotional reactions shared on-line during the Games, especially about the leading and favorite athletes.

Each study session was individual and consisted of two parts: *exploration* of the system and semi-structured *interview*. One session lasted between 40 and 50 minutes, of which the exploration part took $20-25$, and interview – the rest. We asked our users to follow a "thinking aloud" protocol. We also observed how they interacted with the system and recorded their comments about both the interface and the event. All sessions were audio-recorded and transcribed later for the analysis.

In the exploration part of the study, participants interacted with EmotionWatch to investigate the same given Olympic event. This allowed comparison between different users. We chose the women's gymnastics *Balance Beam Final*, an event with many actions throughout its duration, each one triggering emotional responses. The event was also adequate for the comparison view: users would analyze differences in reactions towards the two participating US athletes and athletes from other countries. The video was included. While the event lasted 33 minutes, users were supposed to rather quickly explore its main and most emotional parts.

First, participants were asked to interact with the detailed view interface. After an introduction to the core elements and functionalities, users were then given specific tasks involving working with every element of the interface and allowing them to familiarize themselves with the system and its goals. Some tasks they had to fulfill included:

- Find a moment where anger is the dominant emotion. How was this emotion expressed? What caused it?
- Find a moment with a variety of different emotions. Can you perceive this variety by reading the tweets?

After having performed these guided tasks, we gave participants up to 5 minutes to explore the event further. They were asked to announce during the exploration all findings they considered interesting.

In the second half of the exploration study, users continued exploring the event using the comparison view. After explaining how to work with it, users were given specific tasks to understand the interface better, e.g. to find a moment in time where emotions on either side of the interface differed. Then, they were again given up to 5 minutes to investigate the same event further and provide us with additional interesting observations. They were free to use either the detailed view with keyword filtering, or the comparison view to analyze the reactions towards one or several athletes.

Finally, we conducted a semi-structured interview with the participants. We asked users how they perceived the

emotional reactions to the event and whether the emotions visualized were representative of the event's proceedings. We questioned users on how they had worked with the interface and in which situations they would use either the detailed or the comparison view. We also asked them which functionality they found the most useful and what they would like to see included. We guided users to state their general impression of the system. At the end, we were interested in how users had perceived the association between colors and emotions and whether it was remembered. We presented two other color-emotion associations so the users could state their preferences. Moreover, we were interested in whether the users found the fine-grained approach for emotion categorization to be beneficial.

Results

The general user feedback is summarized in the table 1. Below, P1 to P8 denotes the 8 annonymized users. $n/8$ denotes that n users out of 8 had a consensus on that response. Overall, all individuals successfully used EmotionWatch to research the emotional reactions of people who tweeted about the Olympic events. Below is example feedback about the emotional reactions that our participants discovered:

Now there is more pride – the results became clear (P4)

So much anger to Russians... Where is pity? (P8)

People seem to be angry if they perceive that the athlete was not trying hard enough (P5)

During the post-study interviews, all participants stated that the reactions on Twitter quite likely corresponded to what was captured on the Television and could mostly have been anticipated. They discovered many different emotional reactions caused by this Olympic event, especially *Pride*, *Sadness*, *Happiness*, *Anger*. Moreover, all the users stated that they understood how to interact with the interface, and all gave positive feedback regarding their overall impressions of the system: e.g. *"It was pretty cool, I'd like to see how it works with other events, especially with long-term events like political ones." (P1)* We also observed signs of engagement during the interactions – participants sometimes laughed at the jokes within the tweets ($3/8$).

Users also pointed out valuable directions to improve the current design. For example, three users stated that grasping all the available information was initially complicated: they felt they had to watch the video, read the tweets and look at the emotion wheel at the same time. However, with time, they were able to relate the emotional reactions to a cause and see how they were expressed. In another instance, four users mentioned the difficulty in operating with the variety of colors the system use. We discuss this question in further detail in Emotion Representation Analysis section.

Use of Interface Features

Timeline as the Most Important Feature When asked for the most useful interface features, most users ($6/8$) told us they highly appreciated the overview and the information given by the timeline: both the colors and heights of the bars were helping them to find moments for investigation. They looked for several cues:

- *Peaks* on the timeline, indicating high volume of tweets, served as a cue for emotional intense moments ($4/8$)
- The *aggregates* of multi-colored bars on the timeline within a short period was interpreted as an indicator for an interesting and controversial moment ($3/8$)
- Rare emotions with distinct colors (*outliers* in the timeline), as well as some emotions which were perceived as more interesting (e.g. *Anger*), were considered as worth investigating further ($1/8$)

Furthermore, two users proposed to improve the timeline by encoding the various emotions into each bar on the timeline. Instead of the bar only showing the dominant emotion, we should show the colors of several main emotions with a corresponding distribution. They said this would add an extra cue for finding interesting moments.

Investigating Emotional Reactions To investigate interesting moments more closely in the detailed view, six users were looking at the emotion wheel to get a glimpse of current emotional reactions, while the rest were mostly using the colored emotion names near each tweet for this goal. The text of the tweets then allowed to understand how these emotions were expressed. All users tried to use both tweets and video to reconstruct the causes of the emotions expressed. However, we observed that users relied more on video for this process, probably because tweets did not contain enough direct statements or descriptions of what was happening; video showed this clearly. We asked participants whether they felt they could reach the same findings about emotional reactions if either tweets or video were missing. Five of them affirmed that videos and tweets were both necessary if the event had not been seen before. We were told that the video provided the context for the flow of emotion, while tweets provide the exact emotional reactions. At the same moment, five out of eight participants stated that using video to find

Positive feedback	- Our system helps users to discover and discuss the emotional reactions	O	8/8
	- The reactions quite closely corresponded to the expected ones	A	8/8
	- Multiple emotions are present and perceived	A	6/8
	- Users have understood how to use the system	A	8/8
	- Positive overall impression	A	8/8
	- Interface involve engaging interaction (users laughed)	O	3/8
Problems	- Multi-tasking in following tweets, video and the visualization	A	3/8
	- Multiple confusing colors to present the variety of emotions	A	4/8

Table 1: Summary of the user feedback about the system. O stands for observations, A – for user interview answers.

the cause of an emotional reaction was obstructed by the delay between a filmed occurrence and the tweets discussing it. Users had to discover this fact and adapt their behavior accordingly. Given a moment in time with emotions of interest, they had to jump to previous time intervals and watch the video in order to confirm their idea of what had provoked that reaction. Two users suggested to ease the investigation of specific emotions by allowing filtering by emotion: the possibility to investigate only specific emotions and find their corresponding reactions.

Use of the Comparison View We observed that the emotion wheel received more attention in the comparison view, where it was used as the main element for discovering differences in emotions between two moments – presumably because of the ability to quickly compare the shapes. In turn, the timeline allowed users to grasp the differences in the overall flow of attitudes towards the athletes and moments when they were mentioned more frequently in the tweets. When asked about the appropriate use for the comparison view, all users suggested to compare the athletes or teams. Some gave more specific cases when comparison would be more interesting: for events with one-to-one competitions, e.g. a tennis match (P8); for the event periods where many different actors were involved at the same time, e.g. the moment when the results are published (P4); for comparing the reactions expressed by people from different countries (P1). For the last one, a fair comparison would require language specific emotion recognition systems and similar use of social media in each targeted language.

Analysis of Emotion Representation

Emotion-Color Association Our post-interview included a variety of questions evaluating the chosen spectrum-based association. We discovered that all the participants remembered colors for the main emotions, such as *Anger* (red) or *Sadness* (blue); and that many users $(4/8)$ could correctly describe the color differences for positive and negative emotions (brighter and warmer for positive, darker and colder for negative, with some exceptions). This indicated their ability to rapidly learn our color allocation. Most users $(5/8)$ also perceived the color allocation to be overall representative for the emotions and their structure. However, three users stated that the colors for some emotions were different from what they had expected, e.g. for *Love* and *Involvement*. Three users also reported difficulty in distinguishing some colors – mostly because of the presence of multiple ones.

In the end of the study, we showed participants the screenshots of two other versions of the interface with alternative color allocations: the first one used only black and white alternating around the wheel, while the second used the color-emotion allocation based on the most commonly accepted associations, with an attempt to preserve a separation between colors, but without regard to the wheel structure. Here, we partially used the Plutchik's color wheel (Plutchik 2001) for basic associations, plus our common-sense. While none of the participants preferred the black and white version, $6/8$ participants preferred the second color allocation over our original spectrum. They told us that it corresponded better to their idea of emotions, e.g. that *Love* is better as pink, *Pride* – as green and *Surprise* – as orange.

Fine-Grained Emotion Categorization In our interview, we asked the users if they would prefer to have a polarity-based visualization. All users stated that they strongly benefited from the additional information that the multi-category model provides, and preferred it over a polarity sentiment model. They said that showing polarity alone would not give enough details. Six participants stated it is important to distinguish between specific emotions. Nevertheless, many users $(6/8)$ suggested reducing the number of emotional categories to a range from 4 to 10 to "make it less complicated". After short discussions of possible changes, we found some agreement only in grouping for a few categories (e.g. *Pleasure* with *Happiness* and *Pity* with *Sadness*) and in removing some rare emotions (e.g. *Shame* or *Guilt*).

Conclusion

In this work, we faced the challenge of designing an interactive tool for summarizing the fine-grained emotional reactions expressed in tweets about a public event – EmotionWatch. We proposed a method for visually representing multi-category emotions in chronological order. We applied it to tweets posted during the 2012 Olympic Games to investigate how users perceive and interpret such a representation.

EmotionWatch differs from related event exploration systems in two aspects. First, most other systems focus on reconstructing the event's structure and sub-events, whereas our system's main focus is on presenting people's emotional reactions to the event. Second, instead of merely summarizing the sentiment polarity of tweets, we aim for a multi-category summary of emotional reactions using a fine-grained set of categories. All the users in our evaluation study found it beneficial to distinguish between separate types of emotions instead of just polarity. The study also showed all users successfully interpreted and understood the 20 emotions presented. However, the majority of participants would prefer to have less emotion categories to operate with, while still preserving certain granularity.

Another novel aspect of our work is the visual representation of the interface itself. We suggest combining a color-coded timeline which allows a static overview of the whole event together with an emotion wheel that dynamically expresses the emotional profile of the chosen moment in time. This technique, supplemented with associated tweets and event video, proved to be useful both for the exploration of an event, and for comparing reactions towards event actors.

One challenge in designing EmotionWatch was the allocation of colors to multiple emotions. Our spectrum-based solution received insightful feedback from users. Even with high number of emotions, they appreciated most the color-emotion associations reflecting their personal culturally influenced expectations. Another important issue for users was to keep the categories separable. We suggest that additional work should be done to find an "ideal" allocation.

Our study indicated that the reactions in sports events could be anticipated. We thus assume that our tool could

be even more useful for studying reactions to events in less predictable domains, e.g. to political debates or product announcements. In future, we would adapt our system into a real-time framework to allow for a real-time investigation of the emotional feedback from the crowd to an open-domain event. Then the inclusion of algorithms for on-the-spot sub-event detection and labeling (Marcus et al. 2011; Chakrabarti and Punera 2011; Zhao, Wickramasuriya, and Vasudevan 2011) would allow presenting a structured chronological history of the event, helping to easier identify and investigate the causes of emotional reactions.

Acknowledgments

We thank all the users participating in our study and the anonymous reviewers for their valuable comments. We also thank Marina Boia for collecting the Twitter data, and the Swiss National Science Foundation for their support.

References

Chakrabarti, D., and Punera, K. 2011. Event summarization using tweets. In *Proc. of ICWSM 2011*, 66–73.

Chambers, J. M. 1983. *Graphical methods for data analysis*. Chapman & Hall statistics series. Wadsworth International Group.

Chen, L.; Chen, G.-C.; Xu, C.-Z.; March, J.; and Benford, S. 2008. EmoPlayer: A media player for video clips with affective annotations. *Interacting with Computers* 20(1):17–28.

De Choudhury, M.; Gamon, M.; and Counts, S. 2012. Happy, nervous or surprised? Classification of human affective states in social media. In *Proc. of ICWSM 2012*, 435–438.

Diakopoulos, N.; Naaman, M.; and Kivran-Swaine, F. 2010. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Proc. of VAST 2010*, 115–122. IEEE.

Ekman, P. 1992. An argument for basic emotions. *Cognition & Emotion* 6(3-4):169–200.

Gregory, M. L.; Chinchor, N.; Whitney, P.; Carter, R.; Hetzler, E.; and Turner, A. 2006. User-directed sentiment analysis: Visualizing the affective content of documents. In *Proc. of the Workshop on Sentiment and Subjectivity in Text*, 23–30. ACL.

Hangal, S.; Lam, M. S.; and Heer, J. 2011. Muse: Reviving memories using email archives. In *Proc. of UIST 2011*, 75–84. ACM.

Kamvar, S. D., and Harris, J. 2011. We feel fine and searching the emotional web. In *Proc. of WSDM 2011*, 117–126. ACM.

Kaya, N., and Epps, H. H. 2004. Relationship between color and emotion: A study of college students. *College Student Journal* 38(3):396.

Kim, S.; Bak, J.; and Oh, A. H. 2012. Do you feel what I feel? Social aspects of emotions in Twitter conversations. In *Proc. of ICWSM 2012*, 495–498.

Liu, H.; Selker, T.; and Lieberman, H. 2003. Visualizing the affective structure of a text document. In *CHI'03 Extended Abstracts*, 740–741. ACM.

Marcus, A.; Bernstein, M. S.; Badar, O.; Karger, D. R.; Madden, S.; and Miller, R. C. 2011. TwitInfo: Aggregating and visualizing microblogs for event exploration. In *Proc. of CHI 2011*, 227–236. ACM.

McDuff, D.; Karlson, A.; Kapoor, A.; Roseway, A.; and Czerwinski, M. 2012. AffectAura: An intelligent system for emotional memory. In *Proc. of CHI 2012*, 849–858. ACM.

Miksch, S., and Schumann, H. 2011. *Visualization of time-oriented data*. Springerverlag London Limited.

Mohammad, S. M., and Turney, P. D. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29(3):436–465.

Mohammad, S. M. 2012. #Emotional tweets. In *Proc. of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, 246–255. ACL.

Neviarouskaya, A.; Prendinger, H.; and Ishizuka, M. 2007. Textual affect sensing for sociable and expressive online communication. In *Affective Computing and Intelligent Interaction*, 218–229. Springer.

Nichols, J.; Mahmud, J.; and Drews, C. 2012. Summarizing sporting events using Twitter. In *Proc. of IUI 2012*, 189–198. ACM.

Parrott, W. 2001. *Emotions in social psychology: Essential readings*. Psychology Press.

Plutchik, R. 2001. The nature of emotions. *American Scientist* 89(4):344–350.

Scherer, K. R. 2005. What are emotions? And how can they be measured? *Social science information* 44(4):695–729.

Shamma, D.; Kennedy, L.; and Churchill, E. 2010. Tweetgeist: Can the Twitter timeline teveal the structure of broadcast events. *CSCW Horizons*.

Sharifi, B.; Hutton, M.-A.; and Kalita, J. K. 2010. Experiments in microblog summarization. In *Proc. of SocialCom 2010*, 49–56. IEEE.

Simmons, D. R. 2011. Colour and emotion. *New directions in colour studies* 395–414.

Sintsova, V.; Musat, C.; and Pu, P. 2013. Fine-grained emotion recognition in Olympic tweets based on human computation. In *Proc. of NAACL-HLT WASSA 2013*, 12–20. ACL.

Strapparava, C., and Valitutti, A. 2004. WordNet Affect: An affective extension of WordNet. In *LREC*, volume 4, 1083–1086.

The Guardian. 2012. London 2012 Olympics data. http://www.theguardian.com/sport/series/london-2012-olympics-data.

Tominski, C.; Abello, J.; and Schumann, H. 2005. Interactive poster: 3d axes-based visualizations for time series data. In *Poster Compendium of InfoVis 2005*. Citeseer.

Wang, W.; Chen, L.; Thirunarayan, K.; and Sheth, A. P. 2012. Harnessing twitter "big data" for automatic emotion identification. In *Proc. of SocialCom 2012*, 587–592. IEEE.

Zhao, L.; Wickramasuriya, J.; and Vasudevan, V. 2011. Analyzing Twitter for social tv: Sentiment extraction for sports. In *Proc. EuroITV-FutureTV 2011*.