Deriving Insights from National Happiness Indices

Anthony Brew, Derek Greene, Daniel Archambault, Pádraig Cunningham
Clique Research Cluster, School of Computer Science & Informatics,
University College Dublin, Ireland

***Abstract*—In online social media, individuals produce vast amounts of content which in effect "instruments" the world around us. Users on sites such as Twitter are publicly broadcasting status updates that provide an indication of their mood at a given moment in time, often accompanied by geolocation information. A number of strategies exist to aggregate such content to produce sentiment scores in order to build a "happiness index". In this paper, we describe such a system based on Twitter that maintains a happiness index for nine US cities. The main contribution of this paper is a companion system called SentireCrowds that allows us to identify the underlying causes behind shifts in sentiment. This ability to analyze the components of the sentiment signal highlights a number of problems. It shows that sentiment scoring on social media data without considering context is difficult. More importantly, it highlights cases where sentiment scoring methods are susceptible to unexpected shifts due to noise and trending memes.**

***Keywords*-sentiment analysis; social media; visualization**

I. Introduction

Interest in national happiness indicators stem back as far as 1972 when the king of Bhutan suggested that "Gross National Happiness" was more important than Gross National Product[1]. More recently, the British prime minister has employed Britain's Office of National statistics to develop such metrics[2]. Kramer [1] reported a method to measure national happiness by tracking the word usage in status updates on Facebook. This line of research has received considerable recent attention, with researchers from different disciplines examining the potential to quantify public sentiment by algorithmic analysis of sources such as Twitter and Facebook [2], [3], [4].

In this work, we present a system that uses Twitter to track sentiment across nine US cities. An example of an aggregated index from this system is shown in Fig. 1. To scale to the volumes of data generated by Twitter, we employ a variation of the simple term counting strategy used by Kramer [1] for reasons of interpretability and scale.

The operation of the sentiment tracking index is described in detail in the next section, and a macro-analysis of the index for March–May 2011 is presented in Section IV. This analysis indicates that further tools are required to explain sudden changes in the sentiment signal. Therefore, in section V we describe a companion system, *SentireCrowds*, that allows us to drill down into the data to identify explanations for changes in user sentiment. This system has two core elements: (1) a clustering algorithm for grouping Twitter users based on their tweets in a given time period, and for assigning sentiment scores to these clusters; (2) a visualization tool to support the exploration of topic and sentiment signals over time.

The results of an initial analysis of a large Twitter corpus using this system are presented in Section VI. This analysis highlights some of the challenges and drawbacks of such an algorithmic strategy for sentiment tracking. Our micro-analysis shows that many tweets, that at first might appear to convey sentiment, are simply outbursts of expletives. Presumably, these are an integral part of the signal and must be considered, but they mask the more interesting changes that we seek to identify. In general, our analysis shows that, in sentiment terms, Twitter is a highly-noisy signal. Thus, when designing a sentiment tracking algorithm, it is difficult to avoid filtering and weighting decisions that may bias the index.

II. Related Work

A. Microblogging Data Analysis

Microblogging services allow users to share content by posting frequent, short text updates. Of these services, Twitter has been by far the most popular – expanding rapidly from 94k users in April 2007 [5] to over 200 million unique users by August 2011, with over 200 million posts or "tweets" generated per day[3]. Users can track the content generated by other users based on non-reciprocal "follower" relations.

Many researchers have become interested in exploring content diffusion within the Twitter network, given the potential for Twitter to facilitate the rapid spread of information. Java *et al.* [5] provided an initial analysis of the early growth of the network, and also performed a small-scale evaluation that indicated the presence of distinct Twitter user communities, where the members share common interests as reflected by the terms appearing in their tweets. Kwak *et al.* [6] performed an evaluation based on a sample of 41.7 million users and 106 million tweets from a network mining perspective. The authors studied aspects such as: identifying influential users, information diffusion, and trending topics. Shamma *et al.* [7] performed an analysis on microblogging activity during the 2008 US Presidential Debates. Unlike other text mining tasks, the authors noted that the informal and inconsistent use of vocabulary on Twitter made topic identification difficult. This is exacerbated by the 140 character limit for tweets.

Recently, a variety of researchers have considered Twitter as a target for applying sentiment analysis and opinion mining techniques. Pak & Paroubek [8] collected Twitter data for

[1] http://grossnationalhappiness.com
[2] http://gu.com/p/2y4qc/tw
[3] http://blog.twitter.com/2011/08/your-world-more-connected.html

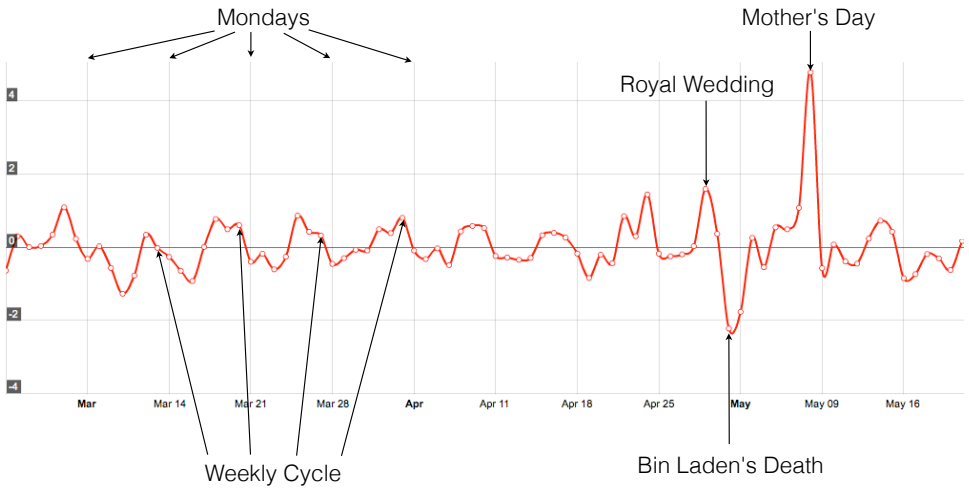


Fig. 1. A plot of the sentiment index for March–May 2011, generated by tracking sentiment on Twitter based on tweets from users in nine cities in the United States. Significant events and periodic effects, corresponding to sentiment peaks and troughs, have been manually annotated.

this purpose and trained a Naïve Bayes classifier on both n-grams and part-of-speech tags to identify positive and negative tweets. Davidov *et al.* [9] performed sentiment classification using different types of features, including punctuation, words, and n-grams. Noisy labels for training were selected based on a small number of pre-specified Twitter hashtags and smileys.

In contrast to previous work, in this paper we describe a system that attempts to both identify specific topics and memes being discussed by users in a Twitter stream, while also exploring the sentiment surrounding these topics.

B. Visualization

A number of systems have looked at ways of visualizing temporally-evolving textual data, some of which have been adapted to visualizing Twitter data. ThemeRiver [10] encodes the frequency of terms as horizontal streams that grow and shrink over time. Dörk *et al.* [11] visualizes conversations in Twitter using a ThemeRiver-like approach. Their system scales to data sets of over a million tweets and successfully identified conversations in the data. Lee *et al.* [12] presented a method that characterizes tags and their evolution in terms of frequency, by overlaying spark lines on each tag. This approach could be applied to visualize trending terms, or sentiment, on Twitter on a per-tag basis.

In a similar way, a number of systems have looked at visualizing document clusters and how they evolve over time. IN-SPIRE [13] creates landscapes of documents using dimensionality reduction based on document statistics. Hetzler *et al.* [14] use animation to depict dynamically evolving clusters and their system has facilities to take snapshots of the data over time. Shi *et al.* [15] combine trend graphs with tag clouds to visualize cluster content and size as it evolves over time.

The above systems are able to visualize changing topics and even changing clusters of text documents, which are similar to Twitter user profiles. However, they are unable to visualize the evolution of clusters at multiple levels of resolution. In this work, we build on the ThemeCrowds [16] system which has been designed to help visualize the tweets of groups of users at an appropriate resolution, along with the evolution of their content over time. In ThemeCrowds, all the tweets a user posts on a given day are stored in a file and a multilevel hierarchy is constructed based on the similarity of those files for each day. By searching or matching clusters via cosine similarity, the system is able to illustrate topics that users are discussing and how the language around those topics changes over time. However, ThemeCrowds is not able to visualize sentiment in conjunction with these topics. In this work, we use sentiment, instead of topic matching, to determine the appropriate level of resolution in the multilevel hierarchies on a given day.

III. A Twitter Happiness Index

Our system for maintaining a Twitter happiness index has been collecting tweets generated by users located in nine US cities: Boston, Chicago, Houston, Los Angeles, Miami, New York, San Francisco, Dallas Fort Worth and Philadelphia. We gathered this data via the Twitter streaming search API by supplying geographic coordinates for one degree by one degree latitude and longitude bounding boxes placed over the center of each city. For the remainder of this paper we focus on a corpus corresponding to all tweets collected between 1 March 2011 to 21 May 21 2011, which consists of 12,781,243 tweets from 336,802 unique users.

We employ the simple term counting strategy presented by Kramer [1] for reasons of interpretability. For each day d, we calculate an overall sentiment score H_d and it is this aggregate daily score that is tracked over time as shown in Fig. 1. The value of H_d is based on the "word count" procedure described in [1]. This approach uses a lexicon of sentiment terms that are associated with positive and negative emotions, and maintains counts of the occurrences of these terms in the tweets collected for a given day. In our system we use a lexicon containing 507 positive and 603 negative terms, which combines a subset of strongly-weighted terms from the

Dictionary of Affect in Language [17] with a manually-curated set of terms and smileys that frequently co-occur in tweets with sentiment terms from the affect dictionary, such as *:)*, *:(*, and *"smh"* (shake my head). Days that result in a higher positive count than average are considered to be positive, while days that contain more negative entries from the lexicon than average are considered to be negative.

Specifically, two scores are calculated for each tweet: a positivity score (percentage of terms that were positive) and a corresponding negativity score. The tweet *"it's a great day"* would get a positivity rating of 0.25 (the term *"great"* is positive while no others are) and a negativity score of 0, while an update of *"super excited :)"* would receive a positivity rating of 1.0 (as all terms in the tweet are considered positive) and a negativity score of 0. In contrast, for a tweet *"It's either good or bad"*, both scores will be 0.2.

These scores are not directly comparable as the usage pattens of sentiment are different, and also due to the fact that our lexicon is unbalanced with respect to the numbers of positive and negative terms. Therefore, we use the formula proposed by Kramer [1] to calculate a normalized "happiness" score for each day:

$$H_d = \frac{\mu_{pd} - \mu_p}{\sigma_p} - \frac{\mu_{nd} - \mu_n}{\sigma_n} \quad (1)$$

where μ_{id} represents the percent of terms that were positive $(i = p)$ or negative $(i = n)$ for a given day d, averaged across every tweet collected. μ_p and μ_n are the overall daily averages and σ_p and σ_n are the standard deviations of across all days analyzed. This approach allows the daily positivity and negativity scores to be normalized so that each contributes in a balanced way to the day's overall happiness score. Thus, a rise in the happiness score may not only be due to increased positive term usage, but could also be due to a drop in negative term usage. An example of this effect is shown in the peak found on Valentines Day, shown in Fig. 2.

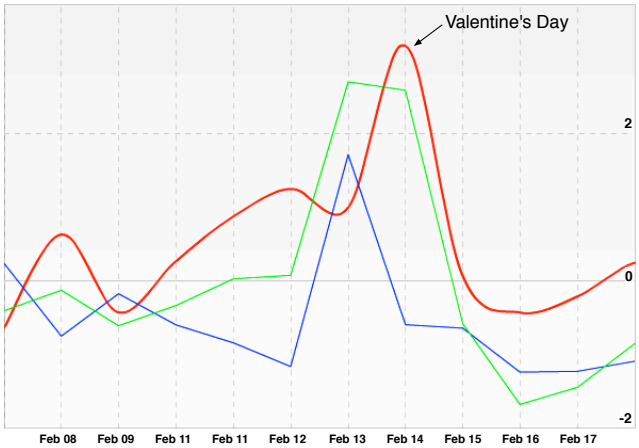


Fig. 2. The "happiness score" may rise not only because of increased positivity, but also because of a drop in negativity, an example of this is the peak found on Valentine's Day (14th February). The positive signal is indicated by the green line, the negative signal by the blue line, and the aggregated happiness score is given by the red line.

Our system demonstrates the viability of building such an index based on tweets, a question raised by Kramer [1] in his discussion on a similar system for Facebook.

IV. Macro-analysis

Before examining the significant peaks and troughs evident in Fig. 1, it is worth looking at the underlying rhythm of the signal. Kramer noted a significant weekly cycle in the Facebook sentiment, and we have observed that the same cycle is evident in Twitter. There is an increase in sentiment over the weekend and a decrease in sentiment after weekends. This can be empirically tested by performing a Fourier decomposition of the signal. This analysis reveals that there is a significant seven-day periodicity in the signal. For certain tasks it may be appropriate to remove this periodicity (*e.g.* to compare sentiment on the same day of the week across multiple weeks).

We now look at the main (non-periodic) peaks and troughs in our data (see Fig. 1). It is interesting that these correspond closely with significant events that occurred during the period covered by the study. For example, the happiest Friday (29 April 2011) coincides with the British Royal Wedding, while the happiest Sunday (8 May 2011) coincides with Mother's Day in the United States. Conversely, the most negative Thursday and Friday (10 and 11 March 2011) respectively coincide with the news breaking about the Japanese earthquake and the subsequent tsunami.

There are however some anomalies in the signal. The death of Osama bin Laden occurred on the 1 May 2011 and shows up as the most negative event in the corpus. This may appear a little surprising given that the tweets originate from the United States – until we see that this trough is due to the increased usage of words such as *"death"* and *"killed"*, which would be naturally annotated as negative in most sentiment lexicons.

We now examine three events in more detail: the Royal Wedding, Bin Laden's Death, and Mother's Day.

A. Top Terms

A simple but effective way to get an overview of the signal is to look at the most commonly used terms for a period after stop-word removal. However, it is clear from Table I that the most frequent sentiment words for a given day are not very informative and appear to contain words/tags that might even be considered stop-words on Twitter. Raw counts do provide some information regarding persistently frequent terms, such as *"lol"* (laugh out loud) and Foursquare geolocation-related tweets (parts of addresses, such as *"st"* and *"ave"*). However, these tend to drown out shifts in sentiment and topical trends. While an analysis of frequent terms indicates what is happening on the days around Bin Laden's death and Mother's Day because the signal is so strong, no relevant topical terms feature prominently on the day of the Royal Wedding.

To find more informative tags, we have sought to identify terms that are *discriminating* for a given day, when compared with previous days. These tags are found by comparing a given day to the proceeding seven days as a baseline. We construct a single aggregated document for each day, calculate

Event	Raw Frequency	Increased Usage
Royal Wedding	lol st like not good love th ave #jobs go	#royalwedding wedding royal #ff st kate ave prince friday york
Bin Laden's Death	lol like not osama st laden good go dead love	osama laden dead obama usa news killed death us president
Mother's Day	lol mothers happy like not love st :) good go	mothers happy mom lakers #ifyoumarryme moms love mother #happymothersday :)

TABLE I
TOP 10 TERMS, AS SELECTED BASED ON RAW FREQUENCY OF OCCURRENCE AND ABOVE-AVERAGE OCCURRENCE, FOR TWITTER DATA COLLECTED ON DAYS CORRESPONDING TO THREE SIGNIFICANT EVENTS.

Event	Sentiment-Associated Terms	Term-Sentiment Bigrams
Royal Wedding	royal #royalwedding #ff kate prince william dress friday watching #icantstandpeoplethat	royal-wedding watching-wedding watch-wedding #royalwedding-wedding #royalwedding-like kate-wedding friday-happy fri-accident #royalwedding-not royal-not
Bin Laden's Death	osama laden obama news president america usa sunday right #ileftyoubecause	laden-dead osama-dead osama-killed laden-killed laden-us osama-us obama-dead osama-death god-bless osama-not
Mother's Day	mom lakers #ifyoumarryme #happymothersday #factsaboutmymom brunch shes mavs church kobe	mom-mothers mom-happy mom-love brunch-mothers dinner-mothers world-mothers one-mothers mom-:) world-happy family-mothers

TABLE II
A BIGRAM-BASED ANALYSIS OF SENTIMENT TERMS FOR DAYS CORRESPONDING TO THREE SIGNIFICANT EVENTS. THE SECOND COLUMN SHOWS TERMS MOST COMMONLY ASSOCIATED WITH SENTIMENT-BEARING TERMS. THE THIRD COLUMN SHOWS FREQUENTLY OCCURRING BIGRAMS THAT CONTAIN A SENTIMENT-BEARING TERM.

a term vector for this document, and normalize it to unit length (as there are varying numbers of tweets found on each day). We then rank the terms based on their increased weight as measured in this normalized representation. It is apparent from the third column in Table I that this strategy highlights informative terms for the days analyzed.

B. Sentiment Co-occurrence

We extend the above analysis by looking at terms that co-occurred with terms in our sentiment lexicon. For each day, we build a document where terms were added to the document if they co-occurred in a tweet with a word in the word lists. We tried two versions: a simple version where the count for a term was incremented each time it co-occurred with a sentiment word, and a *bigram* version that counted bigrams consisting of sentiment and non-sentiment terms (*e.g. "day-good"*, *"mom-happy"*). For both versions, we produced a ranking based on a comparison against a baseline built from the previous seven days. It appears from Table II that the bigram strategy is particularly effective in identifying topics related to sentiment.

V. SENTIRECROWDS

Given the macroscopic profile of how sentiment evolves over time, we may wish to understand why certain events are negative and positive, and what groups of users are saying about those events. In this section, we present a clustering method and visualization system, SentireCrowds, that is able to help explain the sentiment of groups of Twitter users.

A. User Profile Clustering

The visualization component of SentireCrowds takes a time series of multilevel clusterings of Twitter users as its input – each clustering represents a snapshot of discussions on Twitter for a given *time step* (*e.g.* a 24 hour period). Due to the volume of data produced in microblogging platforms, we propose the use of a scalable multilevel agglomerative clustering algorithm, based on the min-max objective described in [18]. The goal of this algorithm is to produce a truncated binary tree, which captures the hierarchical topic structure in the data. This algorithm allows us to generate cluster trees for sequences of data sets containing up to hundreds of thousands of items. A complete description of the algorithm is provided in [19].

In order to cluster users based on the content of their tweets, we follow the user-centric approach of Hannon *et al.* [20]: for each user, we create a single *user profile* document, constructed from the concatenation of all their tweets in a single time step. The scalable clustering algorithm is then applied to the set of user profiles to generate cluster hierarchies for each time step. To provide an intuitive summary of the content of each cluster, we make use of tag clouds. To identify the set of descriptive tags for the clusters in hierarchies generated by our algorithm, we use a centroid-based *concept decomposition* method as proposed by Dhillon *et al.* [21].

Once a clustering of users profiles has been generated, we require a method to produce micro-level sentiment scores on a per-cluster basis for all clusters in the hierarchy. We apply an approach analogous to the macro-level approach described in Section III. As with clustering, sentiment scoring is done on aggregated user profiles in each time step. For each profile in a cluster, we count the frequency of positive and negative sentiment-bearing terms. These counts are normalized with respect to the mean and standard deviation of the positivity and negativity score, as performed in the macro-level sentiment index. The per-cluster sentiment score is calculated as the average score of all profiles assigned to that cluster. One subtle difference between our proposed micro- and macro-level scoring mechanisms is that, due to the use of an unweighted average, users who tweet often do not contribute more to the overall sentiment of the cluster.

B. Visualization System

SentireCrowds is based on the ThemeCrowds [16] system for tracking what groups of users are saying over time. As ThemeCrowds does not convey sentiment information, we had to make a number of modifications to the system in order to support sentiment analysis. The proposed visualization interface for the modified system is shown in Fig. 3.

Our first modification associates a sentiment score with each cluster in the multilevel hierarchy. This sentiment score is used to color each node of the tree, with more saturated colors indicating stronger sentiment (Fig. 4). Positive sentiment is indicated in tan while negative sentiment is indicated in purple. Neutral sentiment is indicated with white and gradients from positive and negative sentiment fade into this neutral color. ThemeCrowds had the ability to find the appropriate resolution of the multilevel hierarchy relative to a particular term or cluster via an automatic maximal antichain selection method. In SentireCrowds, we modify this capability to compute the appropriate resolution based on strength of sentiment by replacing the match score with the pre-computed sentiment score for each cluster. In the case of our system, the appropriate resolution is defined as coarsest resolution of the hierarchy with the most non-neutral sentiment that subtend a subtree s where all clusters in s have more neutral sentiment scores. This maximal antichain must cut all paths in the hierarchy exactly once. SentireCrowds can also operate in the mode where only clusters of positive or negative sentiment are considered.

Secondly, we replace the scented widget used in the previous system with one that illustrates changes in sentiment rather than closeness of match to a topic or cluster. A close-up of this new widget is shown in Fig. 6. The tan part of the timeline above the midpoint of the widget encodes the positive sentiment (sentiment values greater than zero) and

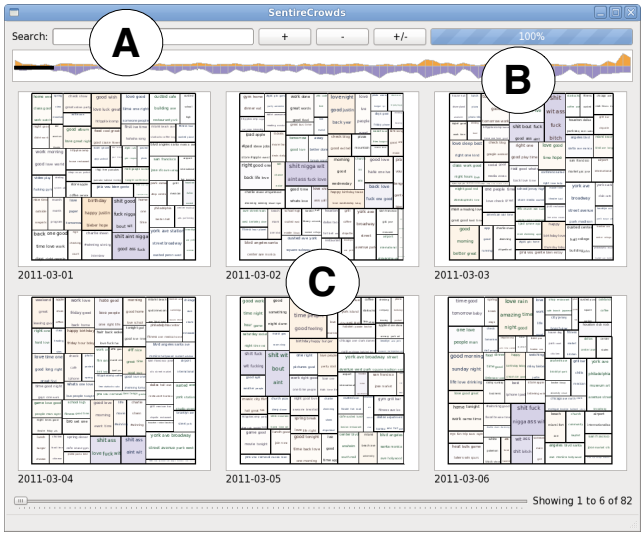
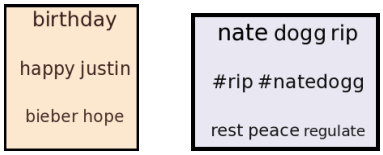


Fig. 3. Main components of the SentireCrowds interface: **(A)** Search box for entering a query term. **(B)** Sentiment widget that depicts the strength of positive and negative sentiment signals over time. **(C)** Small multiples matrix of multilevel tag clouds.



(a) Positive Cluster (b) Negative Cluster

Fig. 4. Positive and negative clusters in the hierarchy are colored tan and purple to indicate their sentiment.

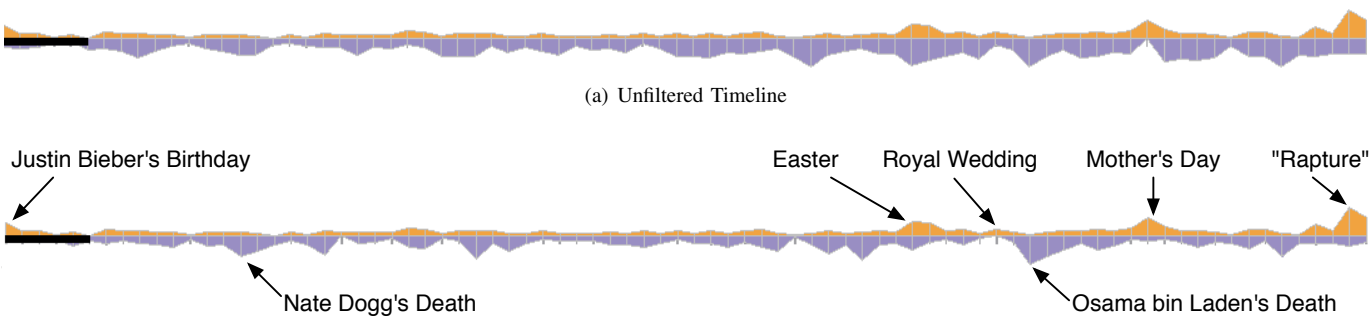
Fig. 5. Sample noisy cluster with strong negative sentiment. These clusters have similar high-frequency terms (in this case mostly expletives), followed by a disparate set of lower-order terms. To interactively filter these clusters from the data set, we look for clusters with similar high-frequency terms and the ignore low-frequency terms.

the purple part of the timeline encodes the negative sentiment. For each day, the heights of the curves represent sentiment score of the most positive and most negative clusters found in the hierarchy.

Under certain circumstances, it can be necessary to remove noisy clusters from the data set in order to see signal in the data. A good example of a noisy cluster is shown in Fig. 5. This cluster contains tweet profiles that are basically swearing clusters. As the clusters contain a large amount of swearing, they are extremely negative and are nearly always the local daily minimum. These clusters can be removed from the sentiment and antichain computations by selecting one of them and *blacklisting* them. The blacklisting operation simply involves determining if a cluster in the hierarchy shares at least a minimum proportion of its ten top tags with the selected cluster – for the results presented here, we use a threshold of $\geq 30\%$ of shared tags.

VI. Micro-Level Analysis

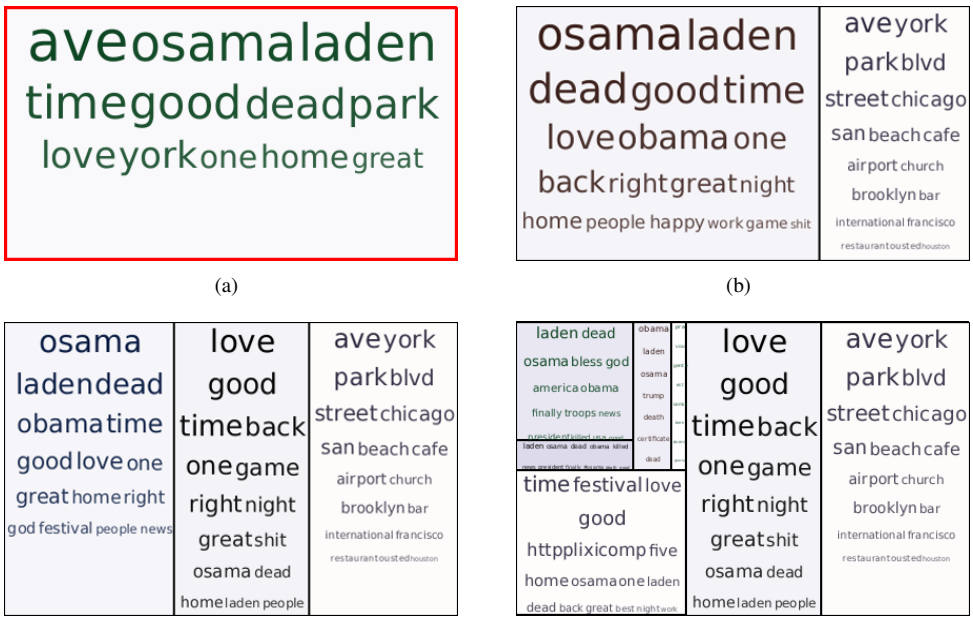
We applied SentireCrowds to the corpus described in Section III. In our analysis we divided this corpus into 82 non-overlapping 24-hour time steps. A multi-lingual stoplist filter was applied to remove non-content-bearing terms – this consisted of 1,937 words from a number of online stoplists in different languages, together with a number of terms commonly appearing in tweets (*e.g.* "RT", "MT"). We also removed Twitter username mentions and URLs. For each time step, we constructed the set of profiles for all users active during that time period – on average each time step contained $\approx 24k$ unique profiles. We applied the scalable min-max agglomerative clustering algorithm to the resulting user profile documents for each 24 hour time step, where the data is divided into $p = 5$ fractions and $k_{low} = k_{high} = 50$ leaf nodes are used to construct comparatively deep hierarchies (see [19] for more details regarding the algorithm parameters).



(a) Unfiltered Timeline

(b) Filtered Timeline

Fig. 6. Timeline in SentireCrowds before and after black listing. Negative sentiment is indicated by purple and positive sentiment by tan. **(a)** The timeline before "expletive" clusters are filtered out of the data. Due to the extreme negative nature of these clusters, they obscure nearly all negative features of the data. **(b)** The timeline after these clusters are removed. Significant events in the negative timeline are now visible, and have been manually annotated.



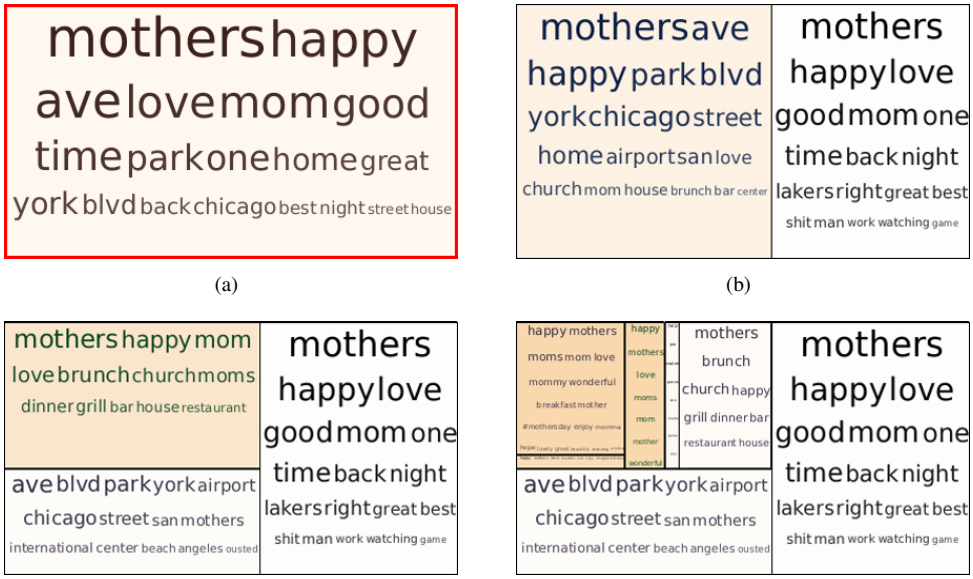
(a) (b)

(c) (d)

Fig. 7. Summaries of tweets in the case study data set on the day of Osama bin Laden's death: **(a)** Root of the hierarchy on the day of his death. **(b)-(d)** Exploration of the hierarchy into deeper levels. Generally, nodes of higher sentiment score tend to be at the leaves of the hierarchy. Using the level of saturation present in the nodes of the treemap, we can follow the sentiment, expanding one node at a time, to the leaves of the hierarchy.

Fig. 6(a) shows the positive and negative SentireCrowds sentiment timelines without any type of filtering applied. Notice that the negative timeline is low for most days and many of the deep troughs correspond to the incidence of clusters where the top tags are expletives, as shown in Fig. 5. We subsequently apply the blacklisting process to remove these noisy clusters to obtain the timeline shown in Fig. 6(b). This timeline illustrates many of the same events shown in the macro-analysis described in Section IV. However, we can now examine which clusters of users are contributing negative or positive sentiment and the topics that they are discussing.

We now take a look at one of the most negative events in the corpus, Osama bin Laden's death (see Fig. 7). Scrolling to the day of Bin Laden's death, we see that at the root of the hierarchy Osama bin Laden is one of the top terms. Drilling one level down below the root (Fig. 7(b)), we see that the content of the user profiles divide more neatly into more neutral Foursquare tweets (*i.e.* tweets generated by the Foursquare social networking system that "checks-in" a Twitter user at a specific geographic location) and slightly negative tweets that contain Osama bin Laden as the top term. Closer to the leaves of the hierarchy, we see the clusters depicted in Figs. 7(c) and 7(d). These levels of the hierarchy have strong sentiment indicated by their more saturated purple color. Looking at the most frequent words that appear in the user profiles of these clusters, we can see stronger sentiment words such as *"dead"*,*"death"*, and *"killed"*. These negative sentiment terms are concentrated in the leaves and less concentrated at higher levels of the hierarchy as they are diluted with more neutral tweets. Thus, by using both the top terms associated with each cluster and following the more negative branches of the hierarchy, we can begin to reason about why particular



(a) (b) (c) (d)

Fig. 8. Summary of the tweets collected on Mother's Day in the United States. This day is the most positive across the entire corpus. **(a)** Root of the hierarchy on Mother's Day. **(b)-(d)** Exploration of the hierarchy towards the leaf clusters.

groups of users are negative and the context of the topic about which they are negative.

At the leaf level (Fig. 7(d)), we see two types of clusters. The largest cluster is a bit more factual in nature, with terms such as *"president"* indicating the announcement of Osama bin Laden's death. The phrase *"God Bless America"* seems to be prevalent in this cluster. The cluster to the right appears to contain users tweeting about Donald Trump's call to release Osama bin Laden's death certificate. Finally, beneath the large cluster, users seem to be simply reporting that Osama bin Laden has died.

In contrast, one of the most positive events in the corpus corresponds to Mother's Day in the United States. At the root of the hierarchy on that day, Fig. 8(a) shows a slightly positive root with *"mothers"* and *"happy"* as the top terms. Drilling down into Figs. 8(b) and 8(c), we see that the sentiment quickly gets concentrated into certain branches of the hierarchy. Fig. 8(d) shows levels of the hierarchy closer to the leaves. We can see three distinct positive clusters. The largest seems to concern itself with wishes but also description of activities with terms such as *"breakfast"* being a frequent term. A second large cluster to the right and a smaller cluster at the bottom, mostly consist of well-wishing tweets. In all cases, terms such as *"happy"*, *"love"*, and *"wonderful"* contribute to the positivity of these clusters.

At the end of the timeline shown in Fig. 6(b), we see a very strong positive spike, starting from May 19th. Initially, this does not appear to correspond to any significant geopolitical, economic, or sporting event around this time. However, using SentireCrowds we can drill down through the cluster hierarchy for the time step corresponding to May 21st to investigate the phenomenon in more detail (see Fig. 9). We see a number of sentiment-bearing clusters that prominently contain the term *"rapture"*. It is apparent that there is significant discussion on Twitter around this date of the prediction by American Christian radio broadcaster Harold Camping that May 21st 2011 would herald "Judgment Day"[4]. The positivity appears to originate from a substantial number of ironic or satirical comments surrounding the story (*e.g. "Pre rapture party. Best idea ever"*, *"I can't think of a rapture joke, I'm not worrying, its not the end of the world"*). It is interesting to note that this meme does not appear to have a significant impact of the sentiment plot shown in Fig. 1. This difference could be due to sentiment scoring using either a per-tweet (as in Fig. 1) or a per-profile basis (as in Fig. 6(b)). The prominence of the signal in the latter suggests that many individual users occasionally tweeted about the topic, while the weak effect in the former suggests that the overall volume of tweets on the topic is relatively low. In fact, only 3.4% of all tweets collected for this day contain the term *"rapture"*. However, 11.8% of the 26,735 user profiles used in clustering contain this term. For profile-based sentiment scoring, the visualization component of SentireCrowds lets us readily explore this unexpected behavior in the sentiment, revealing that the positive spike is due to a trending meme whose popularity does not necessarily reflect the real-world significance of this minor news story.

VII. Conclusion

In this paper, we discussed a system that maintains a happiness index based on sentiment analysis of Twitter. The examples in Section VI illustrated that there is a considerable level of noise in microblogging data. Temporal sentiment analysis in this area is not straightforward, as it is necessary to separate out changes in the signal from the persistent

[4]http://www.bbc.co.uk/news/world-us-canada-13489641

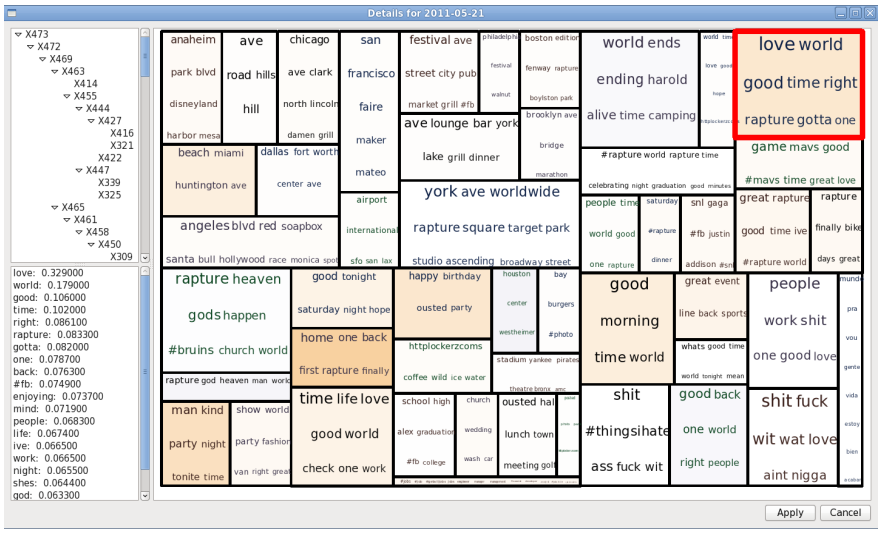


Fig. 9. A screenshot of the SentireCrowds details view, show the cluster hierarchy generated for 21 May 2011. We see a number of sentiment-bearing clusters, mostly positive, discussing the prediction for the "rapture" on this date.

background. We show that a bigram analysis that considers pairs of sentiment-bearing and non-sentiment-bearing terms is effective for indicating the source of sentiment spikes and troughs on a macro level.

However, given that changes in sentiment are associated with specific topics, it is important to be able to organize or cluster social media users and their content, so that we can study sentiment at a higher resolution. We presented SentireCrowds, a multi-resolution clustering and visualization system that supports the exploration of topic-level sentiment in dynamic data. We demonstrated the system on a corpus containing over 12 million tweets, where it highlighted significant events, such as Osama bin Laden's death. The system also helped to explain cases where the sentiment signal shifted unexpectedly, such as when impacted by a trending meme. In the next phase of our work, we plan to conduct a full user study to evaluate the effectiveness of SentireCrowds with users.

Acknowledgment

This work is supported by Science Foundation Ireland Grant No. 08/SRC/I140 (Clique: Graph & Network Analysis Cluster)

References

[1] A. D. Kramer, "An unobtrusive behavioral model of "gross national happiness"," in *Proc. 28th international conference on Human factors in computing systems (CHI '10)*, 2010, pp. 287–290.

[2] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in twitter events," *J. Am. Soc. Information Science and Technology*, 2011.

[3] J. Bollen, B. Goncalves, G. Ruan, and H. Mao, "Happiness is assortative in online social networks," *Artificial Life*, vol. 17, no. 3, pp. 1–15, 2011.

[4] E. Cuvelier and M. Aufaure, "A buzz and e-reputation monitoring tool for twitter based on galois lattices," *Conceptual Structures for Discovering Knowledge*, pp. 91–103, 2011.

[5] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *Proc. Joint 9th WEBKDD and 1st SNA-KDD Workshop*, 2007, pp. 56–65.

[6] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proc.19th international conference on World Wide Web (WWW'10)*. ACM, 2010, pp. 591–600.

[7] D. Shamma, L. Kennedy, and E. Churchill, "Tweet the debates: Understanding community annotation of uncollected sources," in *Proceedings of the 1st SIGMM workshop on Social media*. ACM, 2009, pp. 3–10.

[8] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proc. 7th Conference on International Language Resources and Evaluation (LREC'10)*, 2010.

[9] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in *Proceeding of the 23rd international conference on Computational Linguistics (COLING)*, 2010.

[10] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "ThemeRiver: Visualizing thematic changes in large document collections," *IEEE Trans. on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 9–20, 2002.

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

[12] B. Lee, N. H. Riche, A. K. Karlson, and S. Carpendale, "SparkClouds: Visualizing trends in tag clouds," *IEEE Trans. on Visualization and Computer Graphics (InfoVis 2010)*, vol. 16, no. 6, pp. 1182–1189, 2010.

[13] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, "Visualizing the non-visual: Spatial analysis and interaction with information from text documents," in *Proc. of the IEEE Symp. on Information Visualization*, 1995, pp. 51–58.

[14] E. G. Hetzler, V. L. Crow, D. A. Payne, and A. E. Turner, "Turning the bucket of text into a pipe," in *IEEE Symp. on Information Visualization (InfoVis 2005)*, 2005, pp. 89–94.

[15] L. Shi, F. Wei, S. Liu, L. Tan, X. Lian, and M. X. Zhou, "Understanding text corpora with multiple facets," in *IEEE Symp. on Visual Analytics Science and Technology*, 2010, pp. 99–106.

[16] D. Archambault, D. Greene, P. Cunningham, and N. Hurley, "ThemeCrowds: Multiresolution summaries of twitter usage," in *3rd Int. Workshop on Search and Mining User-generated Content*, 2011.

[17] C. Whissell, "The dictionary of affect in language," *Emotion: Theory, research, and experience*, vol. 4, pp. 113–131, 1989.

[18] C. Ding and X. He, "Cluster merging and splitting in hierarchical clustering algorithms," in *Proc. IEEE International Conference on Data Mining (ICDM'02)*, 2002, pp. 139–146.

[19] D. Archambault, D. Greene, J. Hannon, P. Cunningham, and N. Hurley, "ThemeCrowds: Multiresolution summaries of twitter usage," School of Computer Science & Informatics, UCD, Ireland, Tech. Rep. UCD-CSI-2011-07, June 2011.

[20] J. Hannon, M. Bennett, and B. Smyth, "Recommending twitter users to follow using content and collaborative filtering approaches," in *Proc. 4th ACM Conference on Recommender Systems*, 2010, pp. 199–206.

[21] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine Learning*, vol. 42, no. 1-2, pp. 143–175, 2001.