Under consideration for publication in Knowledge and Information Systems

Event Extraction Using Behaviors of Sentiment Signals and Burst Structure in Social Media

Thin Nguyen[†], Dinh Phung[†], Brett Adams[†] and Svetha Venkatesh[†]

[†]Department of Computing
Curtin University, Perth, Australia.
Email: thin.nguyen@postgrad.curtin.edu.au,{d.phung,b.adams,s.venkatesh}@curtin.edu.au

Abstract. Significant world events often cause the behavioral convergence of the expression of shared sentiment. This paper examines the use of the blogosphere as a framework to study user psychological behaviours, using their sentiment responses as a form of 'sensor' to infer real-world events of importance automatically. We formulate a novel temporal sentiment index function using quantitative measure of the valence value of bearing words in blog posts in which the set of affective bearing words is inspired from psychological research in emotion structure. The annual local minimum and maximum of the proposed sentiment signal function are utilized to extract significant events of the year and corresponding blog posts are further analyzed using topic modelling tools to understand their content. The paper then examines the correlation of topics discovered in relation to world news events reported by the mainstream news service provider, Cable News Network (CNN), and by using the Google search engine. Next, aiming at understanding sentiment at a finer granularity over time, we propose a stochastic burst detection model, extended from the work of Kleinberg, to work incrementally with stream data. The proposed model is then used to extract sentimental bursts occurring within a specific mood label (for example, a burst of observing 'shocked'). The blog posts at those time indices are analyzed to extract topics and these are compared to real-world news events. Our comprehensive set of experiments conducted on a large-scale set of 12 million posts from Livejournal shows that the proposed sentiment index (SI) function coincides well with significant world events while bursts in sentiment allow us to locate finer-grain external world events.

1. Introduction

Social media are a new type of media in which users assume a myriad of roles, including as consumers of information, publishers, editors and commentators. The blogosphere is one example of this new decentralised and collaborative forum, through which people

Received xxx
Revised xxx
Accepted xxx

participate, exchange opinions, generate content and interact with others. This user-generated content mirrors subjective user sentiment, much more so than other written genres. New opportunities exist for opinion mining and sentiment analysis and thus this topic has attracted much recent interest [16, 26, 45]. Sentiment information in social media has also been used to explain real-world trends such as by mapping the proportion of *anxious* posts in Livejournal with the Standard and Poor's 500 index (S&P 500) [25] and predicting the 2009 German election based on political sentiment contained in Twitter [61]. However, joint sentiment–topic investigation, in which sentiment and topic interaction is taken into account together, has received little attention.

Detecting sentiment-based event, Balog et al. [3] found spikes of events based on the mood tags in Livejournal posts. Using a simple threshold approach, they extracted blog posts tagged with bursty moods and used frequent words to interpret related events. Since the bursty intervals were detected based on a simple threshold, a large number were discovered, often occurring over short periods. State–space approaches have been used for burst detection of terms in a document corpus. Kleinberg [33] observes that certain topics in emails are characterised by a sudden increase in textual features. These high-intensity periods of terms, called *burst*, grow in intensity over a period, with the KLB burst detection algorithm being a simple and efficient means of detecting them. However, predefined sets of parameters, obtained from offline data, are required to calculate the probability of generating relevant documents and their state changes. While suitable for offline settings, this method is not scalable for situations in which diverse and changing data is added over time, as in the case of the blogosphere being investigated in this paper.

Therefore, open problems include constructing robust sentiment detection algorithms and incremental burst detection that can be applied to large-scale blog data. Research into emotional responses towards traumatic or significant events has been explored in psychology [2, 6, 54]. Triggered by events in their lives, people write about their experiences in blogs and footprints of their emotional state often emerge in on-line journals. At a larger scale, the blogosphere is a rich source for studying collective emotional response to external events. The question we ask is: Can we use the affect of the blogosphere as indicative of significant world events? Can we use the sentiment in the blogosphere as an indicator of collective emotional states? Can we perform fine-grain analysis on a *specific mood* and examine its correlation to external events?

We examine the sentiment at two levels of granularity—aggregated sentiment and specific sentiment across a corpus. To understand the utility of aggregated sentiment, we estimate the total affective score of all blog posts, constructing a temporal SI. Affective scores of the on-line journals are computed in two ways: first using the valence of mood tagged by the user for each post; and second, by extracting the sentiment-bearing words according to the Linguistic Inquiry and Word Count (LIWC) [46] for 'positive' and 'negative' mood categories. These time-indexed sentiment scores are then analysed to detect maxima and minima, coincident with maximal points in sentiment shifts. The corresponding time indices are used to extract real-world events (from news sites), which are then compared against the topics of the corresponding blog posts, extracted with Latent Dirichlet Allocation (LDA) [5]. In the second part of our work, we examine the extraction of 'bursts' in a particular sentiment across the corpus. We extract time periods of the mood burst based on KLB [33], making adaptations to the algorithm to apply it incrementally to stream data. We extract the topics of the corresponding blogs and examine how the topics correlate with the top news stories extracted from CNN. Our experiments are performed using more than 12 million Livejournal posts and our results demonstrate that the SI is a rich signal, whose maxima and minima correlate well with real-world events. We show that this holds true at a finer level when detecting specific sentiment bursts, and once again, extreme points coincide with finer-grain world events. We demonstrate that the KLB can be efficiently adapted to handle data stream, detecting salient bursty structures in blog data in real-time without compromising performance.

Our contribution in this paper is three-fold. First, we present approaches to construct time-series sentiment indices, whose extrema can be used to correlate to real-world events. We demonstrate that the topics and events extracted correlate well with significant real-world events, validating using the list of top stories voted by CNN. Second, we introduce a novel concept of sentiment burst and employ a stochastic model for detecting bursts in text streams. This provides the foundation for sentiment-based burst detection and subsequently bursty event extraction. Third, we implement an incremental version of the KLB algorithm, which has wide applicability in real-time burst detection in stream data. Evaluating the events extracted is a challenging task. Therefore, an additional contribution is an effective method for evaluating and ranking events extracted using a combination of topic modelling and Google search.

The rest of this paper is organized as follows. Related work to this paper is presented in Section 2. Then, the dataset and evaluation methods are described in Section 3. Next, in Section 4, the sentiment score of the whole sphere in a given time is shown to act as a signal for detecting events discussed among the bloggers. Then, the role of each mood in event detection is studied in Section 5. This section presents the results of bursty pattern discovery of moods and related facts and figures, in different algorithms, and is followed by concluding remarks.

2. Background

Event detection and tracking. Topic Detection and Tracking (TDT) has received much research attention. First sponsored by the Defence Advanced Research Projects Agency in 1996 and then the National Institute of Standards and Technology, TDT-related evaluation tasks have been organised with special interest in 'transcription and understanding of broadcast news'. An emerging part of the TDT initiative is event detection. An *event*, as defined in [1], is something that happens at a certain place and time. Event detection in stream data aims to extract the first story of a new event, while event tracking groups stories discussing the same event. Much work in this topic has followed a content-based approach. For text documents, content features are extracted using unigram or n-grams. Subsequently, a document in a text stream is represented as a vector of features. Event detection and tracking is then cast as a simple nearest-neighbour classification problem based on the similarity of the newly arrived document with the existing pool of events. If the similarity exceeds a predefined threshold, the document is classified into the closest events and updated accordingly to include the new document. Otherwise, a new event is formed.

One of the most popular and useful representations for text is the bag-of-word model, which assumes interchangeability among words in the document. The content represented in this form can be used to answer what happens for an event, either by explicitly examining the words or via topic-modelling tools. Other properties characterising an event, such as *where* the event happens and to *whom* the event is related, can also be estimated through entity recognition [36, 64]. Das Sarma et al. [17] use information about relationships among entities in documents to detect events. Spatial and temporal information can also be exploited to detect context-specific events. Makkonen et al. [39] represent events as vectors of terms and additionally incorporate temporal, location and person identity information. Sakaki et al. [51] utilise spatiotemporal information in Twitter to find the centre and the trajectory of events. Zhao et al. [65] include informa-

Fig. 1. Cloud visualization of 132 predefined mood labels tagged in our Livejournal blog dataset.

tion flow patterns and underlying social network information to detect finer granularity events, where events are defined by keywords (from content-based clustering results) and temporal segments along with community information (based on information flow in a social network). Recently, Subasic and Berendt [57] present a method utilizing local patterns to track the evolution of stories; Tsai and Zhang [60] evaluate the novelty measure of an incoming document based on the novelty of the separate sentences in the document. *However, none of these works has considered the sentiment information as an underlying cause for event detection as presented in this paper*.

Weblogs and sentiment analysis. The blogosphere is growing rapidly. Since bloggers express their feeling and opinions towards certain real life events, the content in blogs tends to be more subjective than in other genres, and thus sentiment analysis to understand social media is growing. Work related to opinion mining and sentiment analysis in social networks can be found in a recent survey by Pang and Lee [45]. Examples for learning affect in the blogosphere include [37], where the authors predict moods for incoming blog posts and detect mood synonymy; or [41], where global mood levels are captured.

One strong form of emotion expression is *mood*. It is a state of the mind such as being happy, sad or angry. It is a complex cognitive process and much debate surrounds its nature, formulation and structure [40, 48, 50]. But better scientific understanding of what constitutes a 'mood' has ramifications beyond psychology alone: for neuroscientists, it might offer insights into the functioning of the human brain; for medical professionals working in the domain of mental health, it might enable better monitoring and intervention for individuals and communities. For instance, one recent study finds empirical evidence for the spread, and influence, of mood among friends and cohorts [11], and has generated tremendous interest among sociologists.[1] Our previous work [44] presents a comprehensive study of different feature selection schemes in machine learning for the problem of mood classification in weblogs. We have also presented data-driven clustering on a similar dataset of over 17 million blog posts with mood

[1] Noticeable interesting results state that one is 15% more likely to be happy if directly connected to another happy person; however if that person is unhappy then the likelihood of happiness decreases by 7%; it also shows that happy and unhappy people tend to cluster themselves.

ground-truth, providing valuable empirical evidence in support of existing psychological models of emotion.

To map moods as well as affective words to their psychological meaning, certain emotion bearing lexicons can be used. This work will use the Affective Norms for English Words (ANEW) [7] to score the affect values of words. ANEW is a set of 1034 sentiment conveying English words, being manually rated in terms of valence, arousal, and dominance a word could convey. ANEW has been used for estimating affect levels in text. For example, Dodds and Danforth [18] use the valence values of ANEW to learn the sentiment degree in three types of data: song lyrics, blogs, and the State of the Union address; Kim and Gilbert [32] use ANEW words to detect sadness in microblogger responses to the death of Michael Jackson in Twitter, finding that the tweets related to the event use many more negative words.

Sentiment information conveyed in social media data can be utilized to characterize blogs. For example, Feng et al. [20] clustered blogs using the hidden sentiment factor. That information is also used in viral marketing, as in Fan and Chang[19], where the authors introduced a sentiment-oriented approach to place advertisements in a specific blog. Sentiment information has also been used to detect significant events. For example, using Livejournal data, Balog et al. [3] found the spike time of events based on the sudden changes in mood–time series. Gilbert and Karahalios [25] estimated the level of '*anxious*' index shown to predict the trend of S&P 500 index. Using Twitter data, Thelwall et al. [59] found that the occurrence of popular events is linked to increases in negative sentiment strength and the index in stock markets. Tumasjan et al. [61] predicted the 2009 German election based on political sentiment contained in Twitter. Using Facebook data, Kramer [34] found that the sentiment contained in status updates varied with occurrences of events. Mishne and Glance [42] utilised blogger sentiment from weblog posts appearing in Blogpulse [27] to predict movie sales.

Behavior Informatics. *Behavior Informatics and Analytics* (BIA) has been recently proposed to support quantitative analysis of behavior [8, 9, 10]. One goals of BIA is to model collective behavior or mass phenomena, such as herding in financial markets [15] and synchronised clapping concert halls [43]. Here we propose an approach to capture the collective sentiment in social media towards events in real-world. There is a causal relationship between emotion and behavior [4]. Thus, our work can be cast as a problem of *pattern analysis of psychological behaviors*, an aspect of BIA [8] that utilizes the collective behavior of people in social networks for the task of event detection and tracking.

Burst detection. Time series data have been extensively studied in data mining. For example, Xing et al. [62] propose ECTS (early classification on time series) to classify early numerical time series data using a minimum prediction length; Saleh and Masseglia [52] use time as a context parameter to discover frequent itemsets. In this study, we focus on a special phenomenon of time series – '*burst*'. As defined in [33], a burst can be understood as an unusual growth in intensity of an observation of interest. A simple approach is to count the relative frequency and apply a threshold for detection. Balog et al. [3] use the thresholding method to detect peak times of moods and explain the spikes by events defined by overused words in the interval. One obvious problem with this approach is the uncertainty in determining a good threshold. Another method for learning peak times for time series is to learn the hidden states generating the observations of the time series. A popular example is the hidden Markov model (HMM) [47]. A seminal work is that by Kleinberg [33], who models the generative process of messages in streams with an infinite state automaton: the higher the state, the higher the rate of generating messages. When the automaton is in high state, a burst is detected at the time interval. Originally, Kleinberg observed that certain topics in his email corpus were more easily characterised by a sudden confluence of message sending, rather than by the textual features of the messages themselves. These high-intensity periods are called bursts, and a term is in a bursty state when it grows in intensity for a period. The Kleinberg model (KLB) is a simple and efficient algorithm to localise a term's burst time.

The KLB has been used to detect events. For example, He et al. [30, 31] use the KLB to detect bursts for unigrams in content and then to explore the co-occurrence of these burst features for event detection. Kumar et al. [35] use the KLB to identify and rank burst in communities by detecting bursty periods of linked structure creation among bloggers in communities. Fujiki et al. [22] modify the Kleinberg model to impose different initial rates on daytime and night-time periods. Gruhl et al. [29] apply the model of detecting bursts to find correlation between on-line content (the blog posts mention) spikes and customer behaviour (sales) spikes.

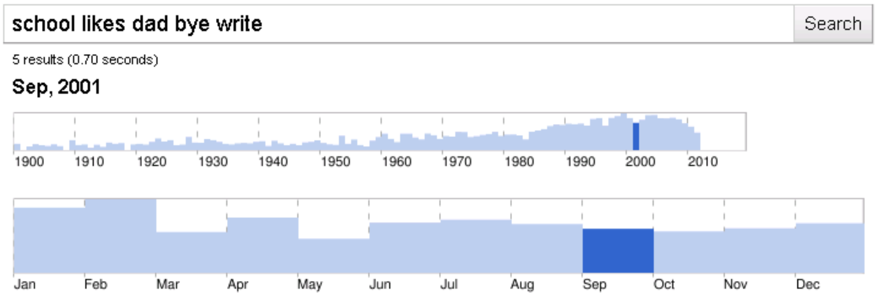
3. Dataset and Evaluation

Dataset. We experiment using data extracted from Livejournal.[2] The dataset was collected for period between 1 May 2001 – 31 December 2004 and contains more than 12 million blog posts tagged with mood labels from a set of 132 moods predefined by Livejournal. Livejournal is a free online blogging service. In addition to conventional posting, it allows users to select one of 132 predefined mood labels to express the user's emotion at the time of writing. This set of predefined moods covers a wide spectrum of emotion, e.g., *cheerful* or *grateful* to reflect happiness, *discontent* or *uncomfortable* for sadness, and so on. Figure 1 shows a tag-cloud visualization of tag-cloud of the moods in our dataset. We note that our method is applicable even when mood tags are not available and thus can be applied to alternative datasets.

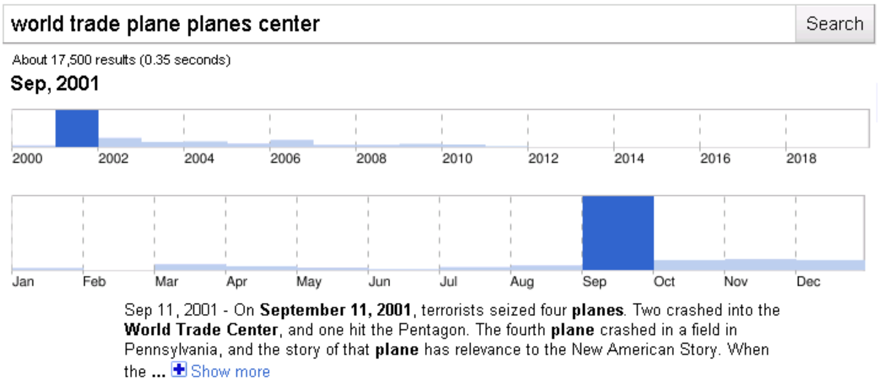
Evaluation method. Evaluation of events detected using our proposed algorithm is challenging due to the scale of the dataset and the unconventional setting of the problem considered in this paper. To the best of our knowledge no benchmark dataset is available. For example, assume our mood burst detection algorithm returns the 1 February 2003 as a significant time index. How do we evaluate if an external event of importance occurred? Two questions remain: *What* is the event? *How* do we know it is significant? One obvious approach is to conduct questionnaires. However, the scope and scale of the dataset quickly rules out the feasibility of this approach.

In this paper, to infer about topics within a post we use a popular Bayesian probabilistic method for topic modelling known as LDA [5]. In the above example, we would collect all blog posts, each of which is a document, posted for the date of 1 February 2003 and run LDA on this corpus, using the returned topics as the content for the posts. To run LDA, we use Gibbs sampling as proposed in [28], which is guaranteed to converge in a limited number of iterations. The topic model requires that the number of topics be specified in advance. A commonly accepted rule of thumb is to choose the number of topics in the scale $\alpha \log V$ where V is the vocabulary size and α is a constant. For our dataset, the vocabulary size is on average equal to $50,000$. A common practice is to use $\alpha = 1$, and using $V = 50,000$, $\alpha \log V \approx 10$. This suggests the number of topics to be 10. Since Gibbs sampling is a special case of Markov Chain Monte Carlo,

[2] http://www.livejournal.com



(a) Timeline search for the top five words for the 'school' topic.



(b) Timeline search for the top five words for the 'WTC' topic.

Fig. 2. Examples of querying Google for top five words for two topics for the date of 11 September 2001. By general search, we cannot determine which topics mention an event since both return millions of results. By restraining the search results to September 2001, 'school' receives five results, whereas 'WTC' returns 17,500 results. Within the timeframe of 10 September–12 September 2001, the search engine returns zero for 'school' and 6,830 results for 'WTC'.

the early samples depend on the initialisation value. To obtain samples which are independent from the initialisation values, it is customary to discard some initial samples as *burn-in*. In our implementation, we discard first $1,000$ samples for *burn-in* and use next $5,000$ samples for inference.

We then propose two approaches to evaluate the significance and impact of the topics returned from our detection algorithms. Firstly, for topics extracted using the SI, we use the set of annual top stories reported by CNN from the corresponding year to evaluate against our extracted topics (see Section 4 for more details). CNN ranking has been chosen as groundtruth for event detection, for example in [12].

However, 'significant event' is an ill-defined term due to its context-specific nature and is often biased by personal knowledge, experience, cultures, and preferences. The CNN ranking is likely biased to a US-centric view of which events qualify as significant, and a populist one at that as the ranking is poll-driven. We therefore supplement the event groundtruth using a method based on Google Timeline, which returns the news articles associated with a given topic within a period of time. Since no polls are involved, this approach offers a better degree of objectivity when compared to the CNN ranking.

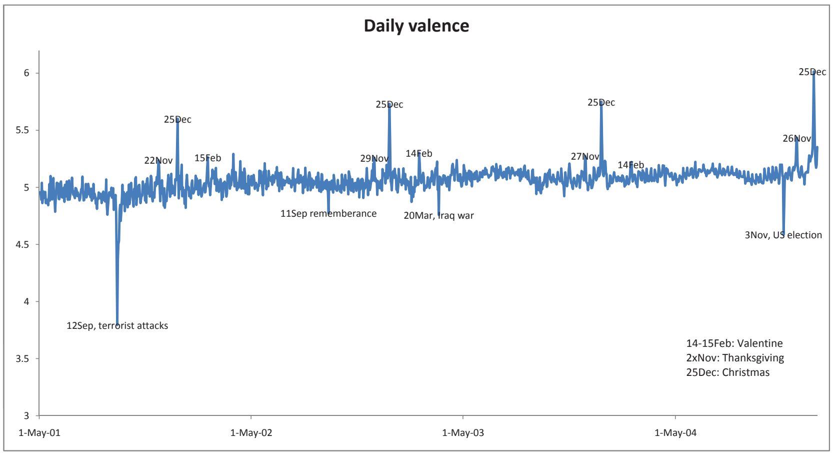


Fig. 3. Sentiment indices computed daily from the whole dataset.

In this second approach, we use the volume returned from Google queried by the top five words per topic during the timespan of the event to examine its significance and impact. To evaluate whether the impact is significant, we construct a baseline measure by randomly selecting five words from the vocabulary set and querying to Google to get the returned volume; we do so 1,000 times and take the average as the reference baseline. An example of this approach is shown in Figure 2. Our intuition is that by restricting the search results to within a limited timeframe (dictated by the time span of the event) the probability of $n = 5$ words simultaneously occurring is extremely unlikely (as validated by the empirical results). Therefore, the impact of events extracted can be justified by the Google results referring to the randomness. We found that this approach is very effective for the task. The average return from Google for the random baseline count is 0.01 ± 0.28, whereas the results from our topics are significantly higher (more details in Section 5). A similar approach has been used in [38], in which the significant level of events is estimated based on the number of documents discussing about it. The more the number of documents reporting about an event, the more significant it is.

4. Sentiment Index and Event Extraction

In this section, we formulate a *temporal SI*. At the global level, sentiment indices serve as a psychological signal, accumulating the average community sentiment across time. The signal provides rich analysis opportunities, for example a large deviation, a local minima or a maxima in this time-series signal could correlate psychological shifts to real-world events. We present two approaches to construct the SI. In the first approach (see Section 4.1), since blog hosting sites (for example, Livejournal) provide a mechanism for the user to tag their mood when the message is posted,[3] we use mood tags and their valence values to construct the SI. However, we realise that these mood tags are not always available. Therefore, in the second approach (see Section 4.2) we construct the SI directly using the affective words used in the blog posts. It is interesting to note at the

[3] That is, the users can indicate if they are sad or happy or angry and so on when they compose the message.

outset that both approaches provide meaningful results that are comparable, suggesting that our sentiment-based methods are applicable in wider contexts.

4.1. Mood-based sentiment index extraction

We describe a method using the emotion information in the mood label tagged to a blog post to discover significant events. Our observation is that a sudden increase or fall in the intensity of emotion expression could be correlated to a real-world event. For example, a flood of *angry* or *shocked* posts is expected following the 9/11 event. To do so, we formulate a novel emotional signal, termed **SI**, capturing the normalised average sentiment over time.

To express mood quantitatively, we need a method to measure the sentimental level of a mood. In the domain of psycholinguistics, one sentiment measure employed by psychologists is *valence*, which indicates the degree of *happiness* a word conveys.

Valence has been found as the principle factor to measure emotion. For example, studies on the structure of English emotion words discover that valence is an important dimension [21, 49, 55]. Rather than English, valence has also been found to be the primary factor to estimate emotions in other languages. For example, Galati et al. [23] discover that the hedonic valence, distinguished by the degree of happiness and sadness, is a major dimension in the organization of the emotional lexicons in Italian, French, Spanish, Catalan, Portuguese and Romanian. Gehm and Scherer [24] analyse typical German emotion words using multi-dimensional scaling (MDS) and find that one fundamental dimension is pleasure/displeasure. Yoshida et al. [63] apply MDS to Japanese emotional words and find that the first dimension is pleasantness-unpleasantness. Shaver et al. [53] argue that different cultures lead to different fine-grained distinctions in the structure of the emotion lexicons (subordinate-level emotion concepts) in American and Indonesian. However, they conclude that emotion concepts in both languages are similar at the superordinate levels, that is, they belong to two categories: positive emotions and negative emotions. Church et al. [13] compare the emotion concepts across cultures, employing samples from English and Filipino. They argue that the nature and range of emotion lexicon of Filipino and English people is similar. In addition, they conclude that the dimensional structure of English emotion terms, in pleasantness and arousal dimensions, is also supported by Filipino emotion words.

However, there is debate in using valence–the degree of pleasure–for measuring of emotions. For example, Colombetti [14] states that pleasures are rarely pure; most of the time they are mixed, that is, they contain pain. Solomon [56] argues that the polarity of positive and negative affects is simple-minded and opposes the idea of defining emotion in terms of their valence.

To alleviate the impact of mixed feelings by individuals, in this paper, for a given mood label, we use the valence value that has been estimated following a consensus basis in ANEW [7].[4] This lexicon contains 1,034 English words, tabled with corresponding sentimental measures of *valence* and *arousal*. These two dimensions have also been used by psychologists in the circumplex model of affect [48, 50] to explain the structure of emotion when emotion states are conceptualised as combinations of these two factors. We then use the valence values as building blocks to compute the SI for the task of event extraction.

Formally, the SI is defined as follows. Denote by $\mathcal{B} = \bigcup_{j=1}^{T} \mathcal{B}_j$ the collection of the

[4] Mood labels which are not listed in the ANEW lexicon are assigned the valence values of their siblings or parents from Livejournal's shallow mood taxonomy (http://www.livejournal.com/moodlist.bml).

Year	Day	Highest ranked topics	Related events	CNN
2001	12 Sep	world trade plane center pentagon planes building towers buildings war york hit school america crashed heard tv watching tower watched	Terrorist attacks	1st
2002	11 Sep	remember september world lost country lives america died 11th american watching bless family live trade plane forget anniversary watch attacks	9/11 anniversary	10th
2003	20 Mar	war iraq bush world country saddam peace president america american support	War in Iraq	1st
2004	03 Nov	bush country years america war president world kerry american voted gay vote election iraq americans rights states marriage hell wrong	2004 election	1st

Table 1. The events detected for the dates of lowest sentiment (based on the topics highest ranked by Google Timeline) and their annual ranking by CNN.

entire dataset where $\mathcal{B}_j$ denotes the set of blog posts arriving for j-th date and T is the total number of days in the corpus. Denote by $\mathcal{M} = \{sad, happy, ...\}$ the set of mood labels predefined by Livejournal. Each blog post $b \in \mathcal{B}$ in the corpus is labelled with a mood $m(b) \in \mathcal{M}$. Denote by v_m the valence value of the mood $m \in \mathcal{M}$. We then compute the SI $I(j)$ for j-th date as:

$$I(j) = \frac{\sum_{b \in \mathcal{B}_j} v_{m(b)}}{|\mathcal{B}_j|} \tag{1}$$

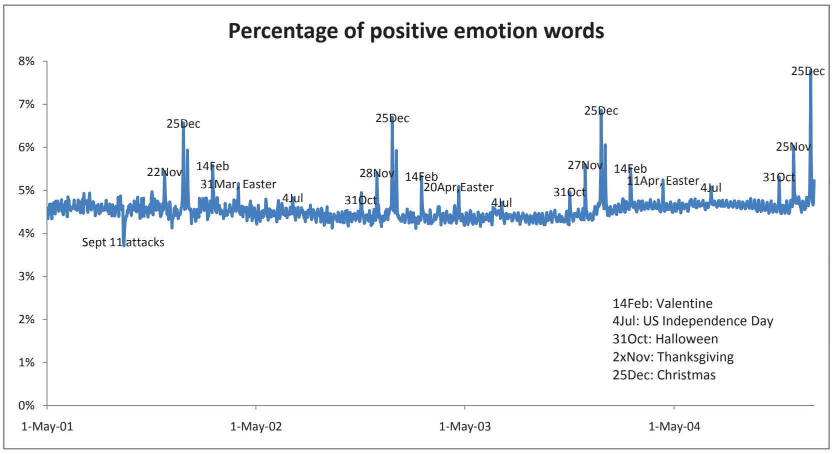
In other words, the SI for a given date is an accumulation of valence values from all posts appearing on that date, normalised by the total count. Figure 3 illustrates the SI computed for the whole dataset.

Examining the structure of the time-series SI $I(j)$ might reveal events of significance. For example, the ebb and flow of this function may be correlated to a psychological shift in the population. Therefore, this high-order sentiment signal provides a fundamental emotion description function and leaves open opportunities for future research. For this paper, in particular, we are interested in the date when the SI reaches its maximum and minimum every year. Figure 3 shows the sentiment indices computed daily. The four highest peaks correspond to highest sentimental dates for 2001 to 2004 respectively; and likewise for the four lowest peaks.

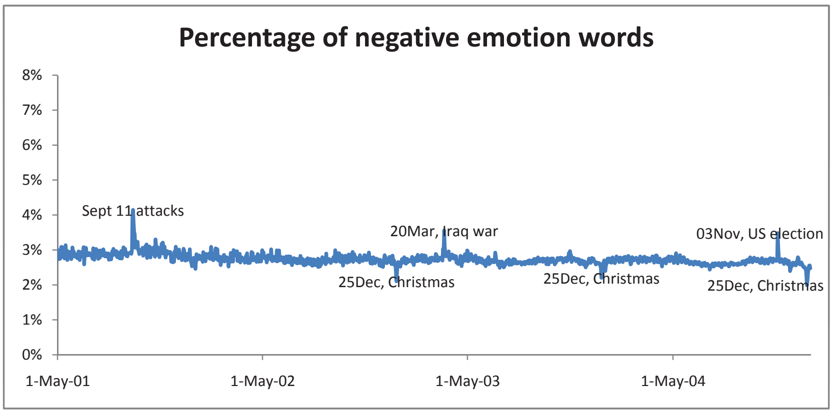
To gain insight into the context of the real-world events correlated with these dates, we retrieve all the blog posts during the corresponding time to form a collection of documents, upon which topic extraction is applied to describe its content. Our results show that the highest valence sentiment at the annual granularity re-occurs on the 25th December—Christmas Day. The highest probability topics learned by LDA from the set of related posts from these dates are dominated by 'Christmas'. These results are quite intuitive, as one would expect many bloggers are '*happy*' during this time of the year. Other public holidays are also found to be high in valence, for example Valentine's, Easter and Thanksgiving (see Figure 3).

In contrast, we find that a majority of the lowest sentiment dates occur during sobering events. For example, in 2002, the lowest sentiment date is 11 September. To learn which topics were discussed on the day, we run LDA over all blog posts made that day (2,973 posts) for ten topics. The list of topics are ranked using the Google method mentioned in Section 3 and are illustrated in Figure 2 with the timeframe of 10 September 2002–12 September 2002. The topic mentioning the 9/11 anniversary scores highest in the Google method by querying 'remember, September, world, lost, country'.

The same process is conducted for other lowest sentiment dates for other years.



(a) Sentiment indices computed daily using *posemo* from the LIWC feature set.



(b) Sentiment indices computed daily using *negemo* from the LIWC feature set.

Fig. 4. Sentiment indices using negative and positive words in blog posts and the corresponding events.

The topics returned from LDA (see Table 1) and ranked highest by the Google method on these dates are compared against the set of top stories ranked by CNN for the corresponding years.[5] The results demonstrate a surprisingly effective set of extracted events: all of the returned events coincide with the first ranked story on the CNN list, with the exception of for 2002. However, even here the detected event is ranked tenth by CNN.

4.2. Content-based sentiment index extraction

Instead of using mood labels, which are not always available, we use sentiment words in the content of the blog post directly to construct our SI. To this end, we utilise words

[5] http://edition.cnn.com/specials/2001/yir/, http://edition.cnn.com/specials/2002/yir/, http://edition.cnn.com/specials/2003/yir/, http://edition.cnn.com/specials/2004/yir/

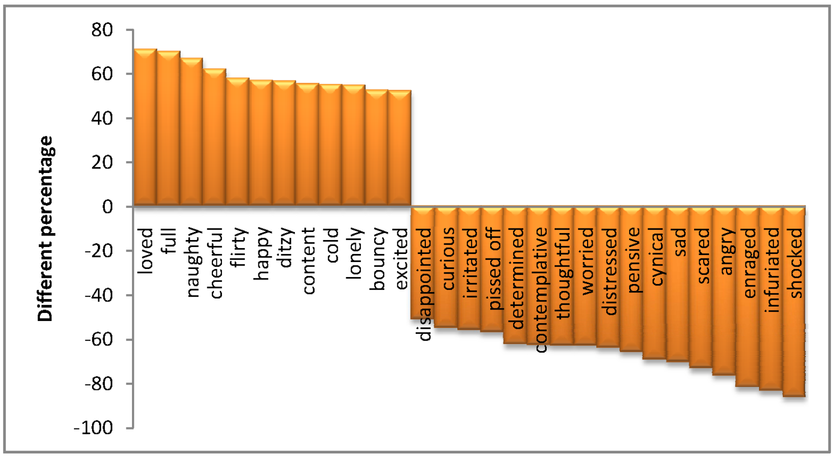


Fig. 5. Happy v. sad events: the moods above the zero line are used more often for happy events and those below the zero line are for traumatic or sad events (***p***s<0.005).

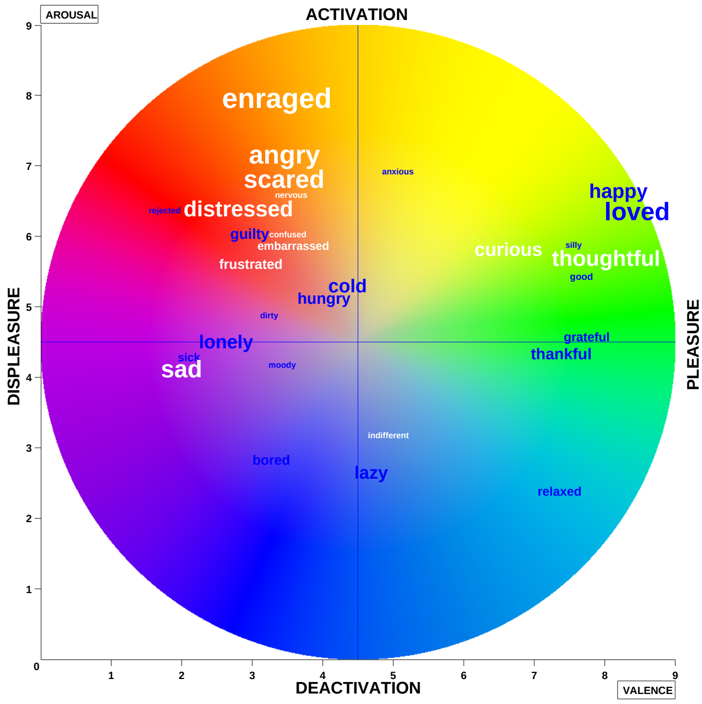


Fig. 6. Happy v. sad events: visualization of corresponding mood labels in the core effect model [48]. Those in blue are used more often for happy events, whereas moods in white are used more often for sobering or sad events. The size of the label reflects the extent of the difference (***p***s<0.005).

from the LIWC [46], which classifies words into a set of linguistic and psychological processes. The LIWC 2007 package and its standard dictionaries [46, 58] are used to extract statistics for the language groups.[6] We then compute the percentage of affective words in a blog post as a measure for sentiment. Two LIWC categories—*posemo* (positive emotion words) and *negemo* (negative emotion words)—are used here as an aggregated indicator of the sentiment reaction of bloggers shown in the content of their blog posts. This is analogous to the two extreme ends of valence values presented earlier: one is for happiness and the other is for sadness.

As expected, as shown in Figure 4a, the highest percentages of positive emotion words in the blog posts are found on public holidays, such as Christmas, Thanksgiving and Valentine's Day, with the lowest percentage of *posemo* words occurring on the anniversary of the September 11 attacks. Conversely, as shown in Figure 4b, *negemo* usage reaches its highest points on the days of traumatic events, such as the September 11 attacks, the Iraq war and the US election, and reaches its lowest points on Christmas Day. This is an almost identical result as found for the mood-based approach, with the exception of the anniversary of September 11 in 2002, suggesting that by exploiting the textual content alone, the proposed method is still effective without the availability of the mood labels.

4.3. Happy Events versus Sad Events

Our observation from the two experiments above is that there is a separation among the mood labels that leads the extraction of events at two extreme ends of emotion, namely *happy* and *sad*, which in turn correspond to the two ends of the valence spectrum. To gain an empirical understanding of how the structure of these moods is distributed, we manually collected 211 topics, in which the top words correspond to happy events, including: '4th July', 'Christmas', 'Easter', 'Halloween', 'Thanksgiving' and 'Valentines'; and 71 topics related to known traumatic events, including 'Iraq war', 'the loss of space shuttle Columbia', 'the attacks of September 11' and 'Ronald Reagan's death'. For this experiment, we run topic model LDA for every day from the corpus and, using the post-topic distributions returned by LDA, we compute the proportion of mood labels for each topic. We then use a standard non-parametric statistical test, Mann–Whitney U, and find that 96 moods are significantly different among happy v. sad events (Mann–Whitney U tests, *p*s < .005 two-tailed, $n_1 = 211$, $n_2 = 71$). Moods with the largest differences are shown in Figure 5; matched with our intuition, *shocked*, *angry*, *sad* and *disappointed* are used more in traumatic events while *happy* and *excited* are used more on holidays. Figure 6 further illustrates the separation of these mood labels using the core affect model proposed by psychologists [48, 50]. We note an interesting result that mood labels used on happy events are found to locate dominantly on the right half of the affect circle, indicating states of *high valence* and *high arousal*. In contrast, for traumatic and sad events, the corresponding mood labels are dominantly *low valence* and *high arousal* and situated in the top-left quarter of the affect circle. A common behaviour of moods in both types of events is that of high arousal, but at two extreme ends of valence.

[6] Descriptions for these groups can be found at http://www.liwc.net/descriptiontable1.php

5. Mood-based Burst and Event Extraction

In the previous sections, we have presented different methods to construct the global SI as a time-series function and events are extracted based on the extreme behaviours of these functions. In this section, we present another approach to extract events based on the burst structure detected with respect to individual mood. Whilst the previous approach characterises aggregated sentiment of the whole population, and is therefore more likely to discover events at a macro level, the burst-based event extraction presented in this section tends to yield more subtle and focused events regulated by the mood label being used. Again, we use the evaluation approach presented in Section 3 to validate performance.

Algorithm 1 Two-state automaton to detect bursts incrementally (iKLB).

for $t = 1$ **to** T **do**
  $\pi_o = R_t/N_t$: non-burst probability.
  $\pi_1 = \lambda\pi_0$: burst probability.
  $\gamma = \ln(t+1)$: transition cost.
  if $(t = 1)$ **then**
    $c_1(s) = -\ln\left[\text{Binomial}(N_1, R_1, \pi_s)\right] + \gamma s$
    $q_1^* = \underset{s}{\text{argmin}} c_1(s)$
  else
    $c_t(s) = -\ln\left[\text{Binomial}(N_t, R_t, \pi_s)\right]$
    if $q_{t-1} = 0$ **then**
      $c_t(s) = c_t(s) + \gamma s$
    end if
    $q_t^* = \underset{s}{\text{argmin}} c_t(s)$
  end if
end for

5.1. Burst-based Event Extraction

Recall our previous notation that the blog dataset is assumed to grow over time $\mathcal{B} = \bigcup_{t=1}^{T} \mathcal{B}_t$ where $\mathcal{B}_t$ denotes the set of blog posts arriving at time t (not necessarily to be within a date).[7] Recall that $\mathcal{M}$ ={*sad*, *happy*,...} the set of moods predefined by Livejournal; each blog post $b \in \mathcal{B}$ in the corpus is labelled with a mood $m(b) \in \mathcal{M}$. The objective is to detect whether there is any burst of posts related to a given mood $m \in \mathcal{M}$ observed over time in $\mathcal{B}$. We appeal to a finite automaton burst detection approach by Kleinberg [33]. Denote by $|\mathcal{B}|$ the number of elements in the set $\mathcal{B}$, the total number of blog posts up to time t is thus:

$$N_t = \sum_{i=1}^{t} |\mathcal{B}_i| = \sum_{i=1}^{t} n_i \tag{2}$$

[7] To be precise, it is a group of blog posts arrived at over time interval t-th, when the time axis is discretized into regular time spans. This is a more realistic scenario in practice than dealing with each individual blog post over time. In the extreme case though, we can still treat each $\mathcal{B}_t$ to contain exactly one blog post.

Score	CNN	Mood – bursty time	Topics	Related events
6390	1	shocked: 11–15Sep01	world trade plane center crashed planes pentagon towers attacks twin york attack president bush terrorist washington united south flight american	September 11
3040	1	angry: 18–20Mar03	war bush country iraq saddam coming america president american support george power fight lives oil hussein stand fear americans	War in Iraq
1750	10	jubilant: 28Oct04	sox red world boston series curse win fans	Red Sox win
909	4	sympathetic: 27–31Dec04	tsunami dead sri death toll earthquake disaster thousands bodies lanka ocean imagine natural thailand goes tidal affected waves countries aid	Natural disasters
527	2	sad: 1Feb03	space shuttle columbia lost challenger crew nasa astronauts families exploded seven watching	Shuttle Columbia
264	1	scared: 30Oct–01Nov04	bush vote kerry election service abortion draft bill voting write agree red children ban supposed health military voted choice country	Election 2004
140	1	stressed: 5–07Oct04	cheney debate edwards country vice second answer vote outside president experience watching listen lack presidential interview enjoy offer involved relax	Election 2004
130	10	sad: 11–12Sep02	lost remember world september lives bless watch watching ones forget hit united stand american real nation died america country heart	9/11 anniversary
71	4	nervous: 30Aug–02Sep04	hurricane frances house hit florida mom friday safe coming deal storm big coast south west family dad computer weather huge	Natural disasters
67	30	scared: 26–31Oct03	fire san fires california live simi homes smoke diego burned school burning firefighters packed state flames house minutes valley miles	California wildfires

Table 2. Top 10 bursty events, according to the Google Timeline results, accompanied by CNN's annual ranking, detected from the bursty time of moods.

where we further use $n_i = |\mathcal{B}_i|$ for brevity. A blog post b is said to be relevant to the mood m if its mood is tagged as m. Therefore, the number of blog posts relevant to mood m arriving at time-slice t, denoted as $r_t(m)$, is

$$r_t(m) = |\{b \mid b \in \mathcal{B}_t, m\,(b) = m\}| \tag{3}$$

Thus, at time t we are given n_t blog posts, and among them there are $r_t\,(m)$ relevant to mood m; thus the total number of blog posts relevant to mood m up to and including time t is:

$$R_t(m) = \sum_{i=1}^{t} r_i(m) \tag{4}$$

A simple approach to detect burst is to plot the histogram of these $R_t\,(m)$ versus n_t and, applying a threshold (THD), declare bursts at those points that the threshold is exceeded [3]. However, as pointed out in [33], THD might easily generate short spurious bursts. [33]'s KLB, a state-based approach to detect bursts, addresses this point. The essential idea of KLB is to model bursting as a generative process using finite automaton. In the simple two-state model, one state is responsible for generating blog posts during non-burst periods and the other is when the burst occurs. The emission probability of the pair $\{R_t, N_t\}$ is modelled a Binomial distribution together with the cost of moving from non-bursty to bursty state – transition cost $\gamma = \ln(T + 1)$. The KLB method reduces to find the optimal sequence of states by minimizing the two costs. The input to the algorithm is an observation sequence of pairs $\{R_t, N_t\}$ for a mood to be detected bursts and the ratio of the emission of relevant documents at state π_0 ($\pi_o = R_T/N_T$) and at state π_1 ($\pi_1 = \lambda\pi_0$). The output is an optimal sequence state $q_1^*, \ldots, q_T^*$ where $q_t^* = 1$ implies burst at time t. We refer readers to [33] for further details and the underlying stochastic explanation for this algorithm.

5.1.1. Incremental burst detection

Since data in the blogosphere is normally large and streams online, a need for a mechanism for detecting bursty periods incrementally for moods arises. While the KLB approach works well in practice, it is not suitable for large-scale and online data since in addition to requiring a multiplier λ to be set manually, this KLB requires advance knowledge of entire data to determine the ratio of emitting relevant documents during non-bursting periods π_0 and the transition cost γ.

We modify the KLB to operate online: the parameter π_o – the rate the automation emitting relevant blog posts during non-bursting periods – is calculated incrementally: that is, π_o is defined as R_t/N_t, where R_t is the number of relevant blog posts until the time slide t and N_t is the total number of blog posts till the time slide t. The transition cost $\gamma = \ln(t+1)$ is also incrementally calculated according to t instead of T. The pseudo code for detecting incrementally bursty intervals of moods is given in Algorithm 1 (iKLB), where Binomial $(n, r, \pi) = (n \text{ chooses } r)\, \pi\, (1-\pi)^{n-r}$ is the usual binomial probability mass function.

5.1.2. Bursty Event Detection

For each mood *m* from the set of 132 predefined mood labels, we perform burst detection using three approaches: THD, KLB and our modification, iKLB. Using the decoded state sequence, a burst is declared when the decoded state transits from low to high. Each bursty intervals of given mood *m* is deemed to be associated with a real-world event; and to infer *what* the event is about, we retrieve the blog posts tagged with *m* during the bursty period and apply topic model LDA to infer the topics, subject to evaluation method presented in Section 3.

5.2. Experimental results

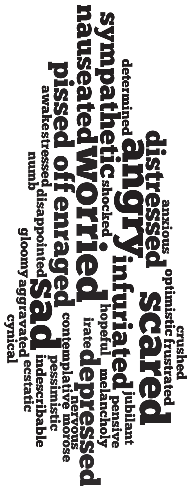
Among 132 predefined mood labels, 76 are bursty (that is, there is at least one burst detected with the corresponding mood label), resulting in 326 intervals. A cloud visualisation of these mood labels is shown in Figure 7a and an on-line demonstration system is provided (see Appendix). We learned topics from blog posts tagged with the bursty mood during the bursty period and then applied Google evaluation (see Section 3) to rank the events and topics. Our detection results are validated as being well correlated with real-world events. For example we find that many events detected are listed in the list of the top 10 CNN stories, as shown in Table 2; all top 10 events detected by our method are in the annual list of top 10 CNN events, except for the last event (however, it is in the list of the top 30 events ranked by the CNN poll).[8]

5.2.1. Inferring impact of mood labels for event extraction

Table 2 shows that high-ranked events tend to be associated with mood labels whose sentiment reflects extreme behaviour, such as *shocked*, *sad* or *angry*. This suggests that the impact of mood labels is not the same when inferring the significance of the derived events. For example, *drunk*, *cold* and *hot* are bursty in many intervals but never found in the top 50 events detected. Using the top 50 events ranked by our Google evaluation method as the reference, we extract 35 moods that are deemed suitable for the task

[8] http://edition.cnn.com/specials/2003/yir/reader.poll.html

(a) The set of bursty moods returned by the burst detection algorithm.



(b) The set of 35 moods found to be significant for event extraction selected from the top 50 ranked events.

(c) Significant moods selected by THD (manually adjusted to obtain 10 bursts per mood label).

Fig. 7. The set of bursty moods (that is, moods with at least one burst detected) and significant moods that are found to be suitable for event extraction (cloud visualisation size is proportional to the number of bursts detected for the respective mood).

of event extraction (see Figure 7b). Among these moods, *angry*, *sad*, *scared* and *worried* are each responsible for three events from the top-50 list; *depressed*, *distressed*, *enraged*, *infuriated*, *nauseated*, *p*ssed off* and *sympathetic* are responsible for a further two events. Consistent with the findings in Section 4, a majority of the moods indicating significant events when bursty are low in valence and high in arousal. High-valence moods—*ecstatic*, *awake* and *jubilant*—are effective at detecting happy events, for example the release of the Harry Potter books or the success of the Red Sox. Likewise, *hopeful* and *optimistic* prove effective for detecting controversial events, for example the US election, 2004.

By comparison, we have applied THD methods with manually adjusted THD so that 10 bursts are detected for each mood label. The results, shown in Figure 7c, are consistent with the findings in the KLB approach: that moods better at indicating significant events are also focused and that a majority of them are low in valence and high in arousal, as shown in Figure 7c. High-valence moods indicate that important events are *ecstatic* (for 'Sox Red win World Series') and *hopeful* (for 'US election').

5.2.2. Top events and burst detection algorithm comparison

To perform a comprehensive evaluation of the proposed algorithms and the quality of the events discovered, we further construct an evaluation dataset on the events as follows: ten peak periods from each mood (with a total of 132) are selected, then corresponding blog posts for each period within each mood are gathered for topic analysis by applying LDA. The number of topics is selected proportional to the size of the vocabulary,

Bursty time	Score	Topics	Related events	CNN rank
26–28Aug01	28	aaliyah plane died family young death fear dead	Aaliyah’s death	30
10–16Sep01	4420	world trade plane center planes crashed york pentagon heard hit buildings second united school building	September 11	1
11–12Sep02	130	lost remember world september lives bless watch watching ones forget hit united stand american real nation died america country heart	9/11 anniversary	10
1Feb03	527	space shuttle columbia lost challenger crew	Shuttle Columbia	2
27Feb03	30	rogers fred died neighborhood mister passed childhood beautiful watching	Mr. Rogers dies	30
11–13Sep03	39	johnny cash died music ritter black	Johnny Cash dies	30
7May04	11	episode end miss rachel watching ross watch watched favorite ended finale goodbye bye cried left joey final cry series season	End of "Friends"	30
11Oct04	38	reeve movie role spinal christopher research sunday cord	Christopher Reeves dies	30
9Dec04	33	dimebag darrell pantera shot metal rip killed guitar band guitarist damageplan	Dimebag Darrell dies	nil
29Dec04	69	tsunami death lost toll red disaster jeg relief asia lives sri earthquake cross world donate thailand thousands aid dead bodies	Natural disasters	4

Table 3. The bursty periods associated with the mood *sad* detected by KLB and the corresponding events, accompanied by the Google Timeline results and annual ranking by CNN. Only the top-ranked topics (by the Google Timeline results) are shown.

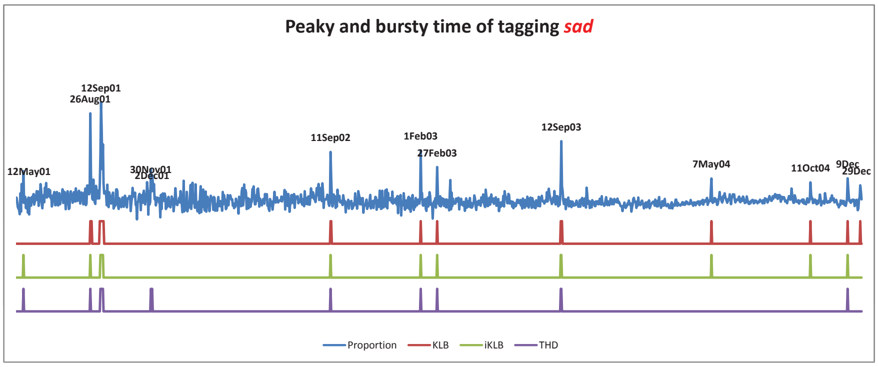


Fig. 8. Highest points (from the normalized signal of the count of blogs tagged with mood *sad*) and burst detection results for mood *sad* from three burst detection algorithms. For THD, we manually adjust the threshold so that ten bursts are detected.

typically 10. For each topic learned by LDA, the top five words and the timestamps (according to the peak period ± 1 day) were submitted to Google Timeline to get the number of pages discussing on the topic within the specified period. We then rank them using the Google volume returns and select the top 70 events as the representative events for our evaluation. In addition, we manually examined each event and labelled the corresponding real-world events. These events and the capacity of both KLB and iKLB to detect them are shown in Tables 4 and 5.

From 326 bursts detected by KLB, we found that 53 intervals match with the time of the 53 top events. Meanwhile, running iKLB on the set of 132 moods resulted in 329

Score	Mood	Peaky Period	Topic words	KLB	iKLB
7740	sympathetic	11-15/Sep/01	world lives trade hit ones	hit	hit
7050	sad	01/Feb/03	shuttle space mission nasa columbia	hit	hit
6370	worried	11-13/Sep/01	world trade planes plane towers	hit	hit
6260	scared	11-16/Sep/01	world trade plane center towers	hit	hit
5410	p*ssed off	03/Nov/04	bush kerry election vote percent	hit	hit
5150	enraged	11-13/Sep/01	world trade center crashed planes	hit	hit
4730	angry	11-16/Sep/01	world trade plane center planes	hit	hit
4370	sad	11-15/Sep/01	world trade plane center planes	hit	hit
4370	shocked	11-15/Sep/01	world trade plane center planes	hit	hit
4230	contemplative	20/Mar/03	war iraq world country live	miss	miss
3880	cynical	03/Nov/04	bush kerry vote election voted	hit	hit
3120	nervous	02-03/Nov/04	bush president war country america	hit	hit
2500	crushed	03-04/Nov/04	bush country kerry vote election	hit	hit
2390	shocked	01/Feb/03	shuttle space nasa crew columbia	miss	hit
1490	disappointed	03-04/Nov/04	war american bush iraq world	hit	hit
1220	disappointed	20/Mar/03	war support bush countries country	miss	miss
1090	enraged	03/Nov/04	election votes florida electoral states	hit	hit
972	pessimistic	03/Nov/04	bush voted country kerry vote	hit	hit
963	angry	03-04/Nov/04	bush country vote voted kerry	hit	hit
883	pessimistic	11-12/Sep/01	center trade started lives nuclear	miss	miss
843	crushed	12-13/Sep/01	america live coming innocent united	hit	hit
842	cynical	20/Mar/03	oil war gets united world	miss	miss
772	grateful	25-26/Nov/04	turkey family thanksgiving food pie	hit	hit
710	grateful	11-12/Sep/02	september events remember lives died	miss	miss
709	angry	20/Mar/03	military international small sites world	hit	miss
699	worried	20/Mar/03	war iraq return world start	hit	hit
288	grateful	27/Nov/03	family thanksgiving food coming big	hit	hit
285	cheerful	22-27/Dec/04	christmas house mom family merry	miss	miss
264	grateful	25-26/Dec/02	couple buy christmas bought wanted	hit	hit
216	thankful	27-28/Nov/03	family thanksgiving especially matter big	hit	hit
203	thankful	28-30/Nov/02	family kind thanks especially light	hit	hit
185	nervous	20/Mar/03	war iraq seeing coming iraqi	miss	hit
176	sympathetic	11/Sep/03	terrorism york left gone lives	miss	hit
174	cheerful	23-26/Dec/01	christmas house mom family eve	miss	miss
163	p*ssed off	11-13/Sep/01	united states stop american act	hit	miss
146	full	27-28/Nov/03	thanksgiving turkey family ate house	hit	hit
130	thankful	25-26/Dec/03	christmas merry family mom gave	hit	hit
120	sympathetic	11-12/Sep/02	knew country died events free	hit	hit
119	scared	20/Mar/03	war world country end away	hit	hit
110	full	25-26/Nov/04	thanksgiving turkey family food eat	hit	hit
109	cheerful	20/Apr/03	easter friday saturday mom gave	miss	miss
107	grateful	25-26/Dec/01	gift gets dad opened bought	hit	hit
102	contemplative	11-12/Sep/02	remember world watched watching hit	hit	hit
102	sad	30/Nov/01	george harrison beatles died beatle	miss	miss
90	thankful	25-26/Dec/02	christmas family gifts lots wants	hit	hit
88	nervous	11/Sep/02	died lives bombings pray afew	miss	miss
84	cheerful	24-26/Dec/03	christmas merry family mom house	hit	hit
70	jubilant	24-25/Dec/04	christmas merry santa holidays holiday	hit	hit
69	hungry	27/Nov/03	thanksgiving eat food family turkey	hit	hit
65	grateful	25-26/Dec/04	christmas family merry mom house	hit	hit

Table 4. The capacity of KLB and iKLB to detect the top 70 events.

Score	Mood	Peaky Period	Topic words	KLB	iKLB
64	grateful	25-26/Dec/03	christmas merry family mom presents	hit	hit
59	thankful	25-26/Dec/04	christmas merry family house mom	hit	hit
56	thankful	24-26/Nov/04	turkey thanksgiving pumpkin potatoes mashed	hit	hit
54	thankful	22-23/Nov/01	thanksgiving family big watching lose	hit	hit
49	sad	11-12/Sep/02	remember world september lives bless	hit	hit
48	hungry	25/Nov/04	turkey thanksgiving eat gobble food	hit	hit
44	cheerful	24-26/Dec/02	christmas merry family presents mom	hit	hit
44	hungry	22/Nov/01	turkey thanksgiving food dinner christmas	miss	hit
40	jubilant	25/Dec/03	letter write personal lyrics christmas	hit	hit
32	sad	12/Sep/03	ritter simple television actor star	hit	hit
30	sad	09/Dec/04	dimebag darrell pantera shot metal	hit	miss
28	jubilant	24-25/Dec/02	started xmas real big beat	hit	miss
28	sad	27/Feb/03	rogers fred died mister neighborhood	hit	hit
25	cheerful	25/Nov/04	thanksgiving turkey family eat gobble	miss	miss
25	full	28-29/Nov/02	thanksgiving holiday holidays celebrated christmas	hit	hit
23	thankful	25/Dec/01	christmas merry family school big	miss	hit
21	shocked	22/Jun/01	hooker lee songs albums rolling	miss	hit
19	full	22-23/Nov/01	turkey thanksgiving eat house dinner	hit	hit
18	grateful	22-23/Nov/01	thanksgiving family words james feels	hit	hit
17	full	26/Dec/04	christmas family merry mom dad	hit	hit

Table 5. The capacity of KLB and iKLB to detect the top 70 events (continued from Table 4).

bursts, among them 55 intervals matched with 55 top events, suggesting a close performance between the offline and the proposed incremental method. The fact that a high proportion (more than 75%) of the top 70 events manually constructed was detected by the algorithm suggests that the proposed scheme is very promising.

5.2.3. Extreme moods and burst detection algorithm comparison

As shown in a previous section (see Figure 7b), *sad* represents an extreme form of emotion and is effective to extract events that are significant. To gain further understanding of the burst structure for this mood label, we present the results from three burst detection algorithms: THD, KLB [33] and iKLB. The results are shown in Figure 8.

We note that KLB performs well on the detection of bursty events. Based on ten bursty periods of *sad* moods detected by KLB, we extract ten corresponding events, extract topics for them and rank these according to the method in Section 3. The results are shown in Figure 3 and we note the significance of the performance given that many of these events are listed in the CNN top stories.

For THD, over ten bursty periods of *sad*, seven events also found by KLB are detected. Three events are missed: the end of the TV series Friends, Christopher Reeve's death and the Indian Ocean earthquake. This method finds an event on 30 November 2001 (the death of George Harrison, lead guitarist of The Beatles) and presents two false alarms on 12 May 2001 and 2 December 2001 when no significant events are found to have occurred. This simple approach works reasonably well on the task. However, it requires advance knowledge of the whole corpus to determine the threshold. For the iKLB algorithm, over ten bursty periods of *sad*, only one event found by KLB (the Indian Ocean earthquake) is missed. This shows that Kleinberg's model can be efficiently adapted as presented in this paper to handle data stream to detect salient burst structures in blog data in real-time without compromising the performance. An on-line demonstration of the bursts and events detected together with their topics and named entities is available (see Appendix).

6. Conclusion

Real-world events often involve shared emotional responses from a large population. We have investigated a novel problem of leveraging the blogosphere as a sentiment 'sensor' to extract events. We have experimented on a large-scale dataset, using millions of blog posts ground truthed with mood labels, and presented a novel method for evaluating the topics and events. We formulated a high-order emotional measure, termed SI, and presented methods for extracting events based on these time-series signals. The results extracted are consistently well matched with lists of top stories as voted by CNN. This demonstrates the usefulness of this sentiment-based approach.

Next, we proposed methods for extracting bursty events by adapting a well-known burst detection algorithm for use on a large-scale dataset. In this setting, data is incrementally added, introducing an on-line implementation for the algorithm. It complements the SI-based approach and the events extracted from the burst detection are specific to individual mood labels at finer levels of analysis of real-world events. After ranking by Google, several of the detected events were found to coincide with the list of CNN top stories.

References

[1] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 37–45, 1998.

[2] M.D. Back, A.C.P. Küfner, and B. Egloff. The emotional timeline of September 11, 2001. *Psychological Science*, 21(10):1417, 2010.

[3] K. Balog, G. Mishne, and M. de Rijke. Why are they excited? Identifying and explaining spikes in blog mood levels. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 207–210, 2006.

[4] R.F. Baumeister, C.N. DeWall, K.D. Vohs, and J.L. Alquist. *Does emotion cause behavior (Apart from making people do stupid, destructive things)?*, chapter 7, page 119. Oxford University Press, 2010.

[5] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] P.J. Bracken, J.E. Giller, and D. Summerfield. Psychological responses to war and atrocity: The limitations of current concepts. *Social Science & Medicine*, 40(8):1073–1082, 1995.

[7] M.M. Bradley and P.J. Lang. Affective norms for English words (ANEW): Instruction manual and affective ratings. *University of Florida*, 1999.

[8] L. Cao. Behavior informatics and analytics: Let behavior talk. In *Proceedings of the IEEE International Conference on Data Mining Workshops*, pages 87–96, 2008.

[9] L. Cao. In-depth behavior understanding and use: The behavior informatics approach. *Information Sciences*, 180(17):3067–3085, 2010.

[10] L. Cao, Y. Ou, and P.S. Yu. Coupled behavior analysis with applications. *IEEE Transactions on Knowledge and Data Engineering*, PP(99):1, 2011.

[11] N.A. Christakis and J.H. Fowler. *Connected: The surprising power of our social networks and how they shape our lives*. Little, Brown and Company, 2009.

[12] A.Y.K. Chua, K. Razikin, and D.H. Goh. Social tags as news event detectors. *Journal of Information Science*, 37(1):3, 2011.

[13] T. Church, M. S. Katigbak, J.A.S. Reyes, and S.M. Jensen. Language and organisation of Filipino emotion concepts: Comparing emotion concepts and dimensions across cultures. *Cognition & Emotion*, 12(1):63–92, 1998.

[14]G. Colombetti. Appraising valence. *Journal of Consciousness Studies, 12*, 8(10):103–126, 2005.
[15]R. Cont and J.P. Bouchaud. Herd behavior and aggregate fluctuations in financial markets. *Macroeconomic Dynamics*, 4(02):170–196, 2000.
[16]R. Coontz. Blogs: Happiness barometers? *Science*, 325:5941, 2009.
[17]A. Das Sarma, A. Jain, and C. Yu. Dynamic relationship and event discovery. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 207–216, 2011.
[18]P.S. Dodds and C.M. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456, 2010.
[19]T.K. Fan and C.H. Chang. Sentiment-oriented contextual advertising. *Knowledge and Information Systems*, 23:321–344, 2010.
[20]S. Feng, D. Wang, G. Yu, W. Gao, and K.F. Wong. Extracting common emotions from blogs based on fine-grained sentiment clustering. *Knowledge and Information Systems*, 27:281–302, 2011.
[21]J.R.J. Fontaine, K.R. Scherer, E.B. Roesch, and P.C. Ellsworth. The world of emotions is not two-dimensional. *Psychological Science*, 18(12):1050, 2007.
[22]T. Fujiki, T. Nanno, Y. Suzuki, and M. Okumura. Identification of bursts in a document stream. In *Proceedings of the First International Workshop on Knowledge Discovery in Data Streams*, pages 55–64, 2004.
[23]D. Galati, B. Sini, C. Tinti, and S. Testa. The lexicon of emotion in the neo-Latin languages. *Social Science Information*, 47(2):205, 2008.
[24]T. Gehm and K. R. Scherer. *Factors determining the dimensions of subjective emotional space*, chapter 5, pages 99–114. Lawrence Erlbaum Associates, 1988.
[25]E. Gilbert and K. Karahalios. Widespread worry and the stock market. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.
[26]J. Giles. Blogs and tweets could predict the future. *The New Scientist*, 206(2765):20–21, 2010.
[27]N. Glance, M. Hurst, and T. Tomokiyo. Blogpulse: Automated trend discovery for weblogs. In *Proceedings of the WWW Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004.
[28]T.L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(90001):5228–5235, 2004.
[29]D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 78–87, 2005.
[30]Q. He, K. Chang, and E.P. Lim. Using burstiness to improve clustering of topics in news streams. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 493–498, 2007.
[31]Q. He, K. Chang, E.P. Lim, and J. Zhang. Bursty feature representation for clustering text streams. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pages 26–28, 2007.
[32]E. Kim, S. Gilbert, MJ Edwards, and E. Graeff. Detecting sadness in 140 characters: Sentiment analysis of mourning Michael Jackson on Twitter. Technical report, Web Ecology Project, 2009.
[33]J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.
[34]A.D.I. Kramer. An unobtrusive behavioral model of gross national happiness. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (SIGCHI)*, pages 287–290, 2010.
[35]R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 568–576, 2003.
[36]G. Kumaran and J. Allan. Text classification and named entities for new event detection. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 297–304, 2004.
[37]G. Leshed and J.J. Kaye. Understanding how bloggers feel: Recognizing affect in blog posts. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (SIGCHI)*, page 1024, 2006.
[38]D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. Keim. Eventriver: Visually exploring text collections with temporal references. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1, 2011.
[39]J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi. Topic detection and tracking with spatio-temporal evidence. *Advances in Information Retrieval*, pages 549–549, 2003.
[40]I.B. Mauss and M.D. Robinson. Measures of emotion: A review. *Cognition & Emotion*, 23(2):209–237, 2009.
[41]G. Mishne and M. De Rijke. Capturing global mood levels using blog posts. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, pages 145–152, 2006.
[42]G. Mishne and N. Glance. Predicting movie sales from blogger sentiment. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.
[43]Z. Néda, E. Ravasz, Y. Brechet, T. Vicsek, and A.L. Barabási. Self-organizing processes: The sound of many hands clapping. *Nature*, 403:849–850, 2000.

[44]T. Nguyen, D. Phung, B. Adams, T. Tran, and S. Venkatesh. Classification and pattern discovery of mood in weblogs. *Advances in Knowledge Discovery and Data Mining*, pages 283–290, 2010.
[45]B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.
[46]J.W. Pennebaker, M.E. Francis, and R.J. Booth. Linguistic inquiry and word count (LIWC) [computer software]. *Austin, Texas: LIWC Inc*, 2007.
[47]L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
[48]J.A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.
[49]J.A. Russell. Pancultural aspects of the human conceptual organization of emotions. *Journal of Personality and Social Psychology*, 45(6):1281, 1983.
[50]J.A. Russell. Emotion, core affect, and psychological construction. *Cognition & Emotion*, 23(7):1259–1283, 2009.
[51]T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 851–860, 2010.
[52]B. Saleh and F. Masseglia. Discovering frequent behaviors: Time is an essential element of the context. *Knowledge and Information Systems*, pages 1–21, 2010.
[53]P.R. Shaver, U. Murdaya, and R.C. Fraley. Structure of the Indonesian emotion lexicon. *Asian Journal of Social Psychology*, 4(3):201–224, 2001.
[54]R.C. Silver, E.A. Holman, D.N. McIntosh, M. Poulin, and V. Gil-Rivas. Nationwide longitudinal study of psychological responses to September 11. *Journal of the American Medical Association*, 288(10):1235, 2002.
[55]C.A. Smith and P.C. Ellsworth. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48(4):813, 1985.
[56]R. C. Solomon. Against valence ('positive and negative emotions'). *Not Passion's Slave*, 1(9):162–178, 2003.
[57]I. Subasic and B. Berendt. Discovery of interactive graphs for understanding and searching time-indexed corpora. *Knowledge and Information Systems*, 23:293–319, 2010.
[58]Y.R. Tausczik and J.W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24, 2010.
[59]M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418, 2011.
[60]F. Tsai and Y. Zhang. D2S: Document-to-sentence framework for novelty detection. *Knowledge and Information Systems*, pages 1–15, 2010.
[61]A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.
[62]Z. Xing, J. Pei, and P. Yu. Early classification on time series. *Knowledge and Information Systems*, pages 1–23, 2011.
[63]M. Yoshida, R. Kinase, J. Kurokawa, and S. Yashiro. Multi-dimensional scaling of emotion. *Japanese Psychological Research*, 12(2):45–61, 1970.
[64]K. Zhang, J. Zi, and L.G. Wu. New event detection based on indexing-tree and named entity. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 215–222, 2007.
[65]Q. Zhao, P. Mitra, and B. Chen. Temporal and information flow based event detection from social text streams. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 1501–1506, 2007.

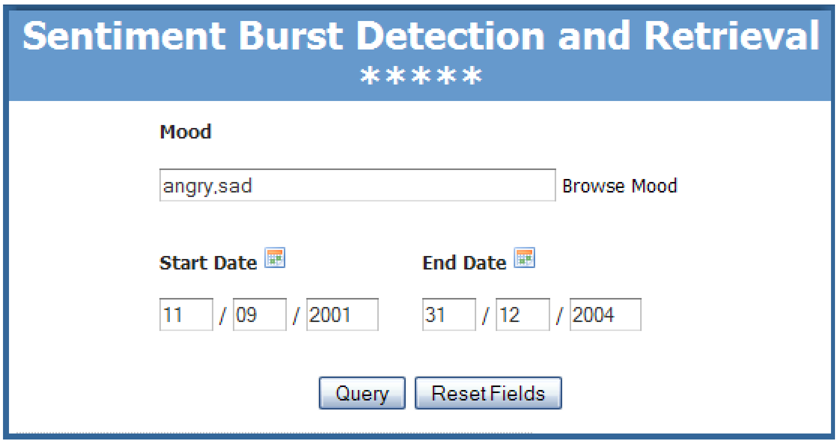


Fig. 9. The input Web user interface for querying bursty moods and related events.

A. Sentiment Burst Detection and Retrieval System

A brief on-line demonstration of the system for burst and events detection based on mood labels can be found at http://blog-research.ivec.org/burst/. Figures 9 and 10 display screen-shots from the Web user interface. A user can enter a set of moods and/or the timeframe to query (9) and the software subsequently returns bursts associated with those moods with timestamp during the bursty duration. For example, a query of '*angry*, *sad*' and duration '11 September 2001 to 31 December 2004' results in 3 and 8 bursty events corresponding to the moods *angry* and *sad* respectively. Figure 10 provides a detailed result including topics and entities for the event '9/11 attacks' and the timeline for two other events associated with the mood *angry*, including 'Iraq war' and 'US election'.

Fig. 10. Partial result of the bursty moods, related events, topics and named entities returned for the query shown in Figure 9.