

Sentiment Index and Bursty Event Detection in the Blogosphere

No Author Given

No Institute Given

Abstract. An important problem in text mining is to detect bursty events. For example, knowing when a topic rises, falls or fades away has important implications in text indexing and retrieval systems, in market and consumer predictions, and advertising strategy. We introduce a *sentiment index*, computed from the *current mood* tags in a collection of blog posts utilizing an affective lexicon, potentially revealing subtle events discussed in the blogosphere. We then develop a method for extracting events based on this index and its distribution. Our second contribution is establishment of a new bursty structure in text streams termed a *sentiment burst*. We employ a stochastic model to detect bursty periods of moods and the events associated. Our results on a dataset of more than 12 million mood-tagged blog posts over a 4-year period have shown that our sentiment-based bursty events are indeed meaningful, in several ways.

1 Introduction

Social media is a new type of media where readers, along with their conventional roles as information consumers, can be publishers, editors, or commentators. The blogosphere, a popular representative of the new decentralized and collaborative media model, is where people can participate, express opinions, mediate their own content, and interact with other users. The user-generated content in social media tends to be more subjective than other written genres. This opens new opportunities for studying novel pattern recognition approaches based on sentiment information such as those initially reported in [3,6].

Research in opinion mining and sentiment analysis, which commonly deals with the subjective and opinionated part of data, has recently attracted a great deal of attention [9]. Sentiment information in social media has been used to explain real world settings, such as mapping the proportion of *anxious* posts in Livejournal with the trend of the S&P 500 index in stock market [5], or predicting the 2009 German election based on political sentiment contained in Twitter [12].

Due to the growing scale of the blogosphere, detecting bursty events in the media is an emerging need. Existing work has looked at simple burst detection of a term in a collection of documents, including the popular state-machine developed by Kleinberg [8] (KLB). Kleinberg observes that certain topics in his emails may be more easily characterized by a sudden increase in message sending, rather than by textual features of the messages themselves, and he calls these high intensity periods of sending *bursts*. A term is in bursty time when it grows in intensity for a period.

2.1 Sentiment index for event detection

Livejournal allows users to tag a new post with their *current mood*. Its set of 132 predefined moods covers a wide spectrum of emotion. Examples are *cheerful* or *grateful* to reflect happiness, *discontent* or *uncomfortable* for sadness, and so on. Figure 1 contains a tag-cloud of moods in our dataset. A sudden increase in tagging of certain moods could be due to a real-world event. E.g., a flurry of *angry* posts might occur following 9/11 event. To detect the trend of mood patterns used in a period, which is termed *sentiment index*, we summarize the emotion quantity of the moods in the period.

One widely accepted emotion measure used by psychologists in text analysis is *valence*, which indicates the level of *happiness* a word conveys. To measure the valence value a mood label conveys, we adapt the affective norms for English words (ANEW) [2], which contains 1034 English words with assigned values for *valence* and *arousal*. These two dimensions are also used in the circumplex model of affect [10,11], where emotion states are conceptualized as combinations of these two factors. We then use the valence values as building blocks to compute the sentiment index for extraction of events. Those moods which are not in the ANEW lexicon are assigned the valences of their siblings or parents from Livejournal’s shallow mood taxonomy².

Formally, the sentiment index is defined as follows. Denote by $\mathcal{B} = \{\mathcal{B}_1, \dots, \mathcal{B}_T\}$ the collection of the blog dataset, where \mathcal{B}_t denotes the set of blog posts arriving at day t , and T is the total days of the corpus. Denote by $\mathcal{M} = \{sad, happy, \dots\}$ the set of moods predefined by Livejournal. Each blogpost $b \in \mathcal{B}$ in the corpus is labeled with a mood $l_b \in \mathcal{M}$. Denote by v_m the valence value of the mood $m \in \mathcal{M}$. We then formulate the sentiment index I_t at the day t^{th} as

$$I_t = \frac{\sum v_{l_b} | b \in \mathcal{B}_t}{|\mathcal{B}_t|}$$

The average sentiment index \bar{I} is computed as

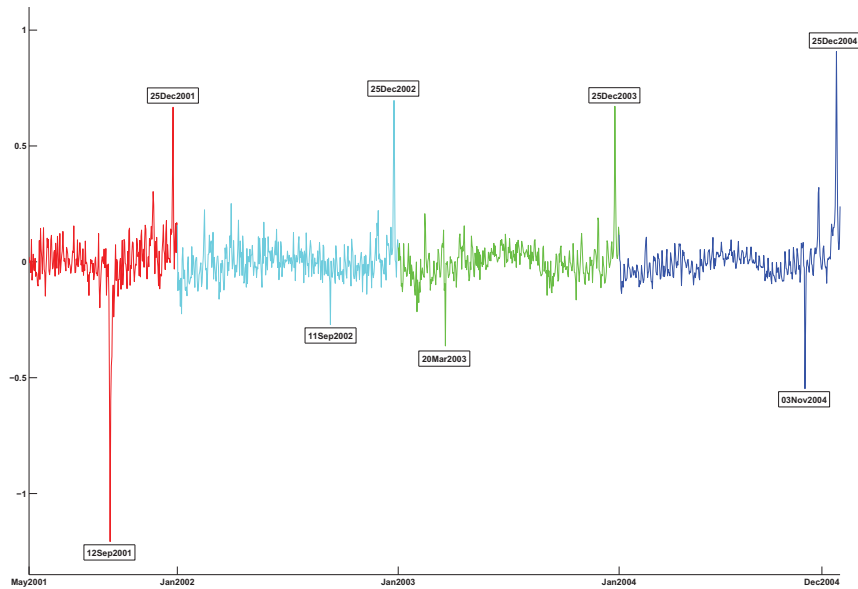
$$\bar{I} = \frac{\sum_{t=1}^T I_t}{T}$$

We define $\delta_t = I_t - \bar{I}$, the sentiment deviation from the sentiment mean for day t^{th} . A plot of this value for the corpus period is shown in Figure 2a.

We conjecture that extreme deviations δ_t can potentially lead to the discovery of important events. Thus, for each year we examine the time when the sentiment index reaches its maximum or minimum. We extract the time when the extremes occur as signalling events to be derived. To get insights into the content of these events, we retrieve all the blog posts during the corresponding periods to form a collection of documents, upon which topic extraction or NLP techniques can be applied to describe its content.

We use Latent Dirichlet Allocation (LDA) [1] to infer latent topics. Since exact inference is intractable for LDA, we use Gibbs sampling, proposed in [7], for learning blogpost–topic distribution. To extract entities corresponding to *person*, *location*, and

² <http://www.livejournal.com/moodlist.bml>



(a) Sentiment deviation.

| Year | Events extracted by our proposed sentiment index | Entities extracted by our proposed sentiment index | CNN reports on top stories |
|------|--|--|---|
| 2001 | | | The terrorist attacks (ranked first) |
| 2002 | | | 9/11 anniversary (ranked tenth) |
| 2003 | | | War in Iraq (ranked first) |
| 2004 | | | 2004 election (ranked first) |

(b) The events detected for the days of lowest sentiment.

Fig. 2: Sentiment deviation and the events detected in the lowest sentiment days.

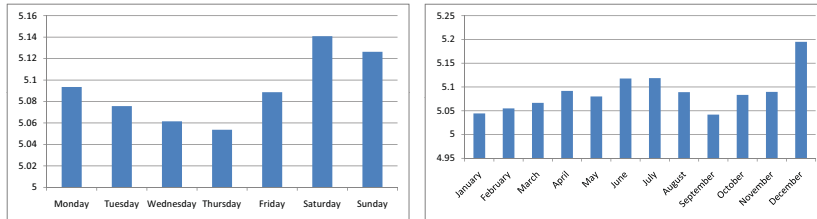


Fig. 3: Weekly and annual patterns of sentiment indices.

organization in the text, we use the Stanford Named Entity Recognizer (Stanford NER) [4].

Our results show that the highest valence sentiment at the annual granularity re-occurs on the 25th December – Christmas Day. The top probability topics learnt by LDA from the set of related posts in these days also dominantly mention Christmas. These results are quite intuitive, as one would expect many bloggers would be ‘happy’ during this period of the year.

In contrast, we find that all the lowest sentiment days are in the time of sobering events. Both topics and entities (Figure 2b) returned from running LDA and Stanford NER on the blog posts on these days mention the top stories reported by CNN³. Notably, all the extracted events are ranked first in the CNN lists, except the event detected in 2002 which is ranked tenth.

To examine periodic events or bloggers’ behaviors, we aggregate distributions of the valence values by weekdays and by months. As shown in Figure 3, the sentiment index is lowest on *Thursday* and highest on *weekends*. When looking at monthly patterns, *December* is top in the index in a month of many celebrated days, such as Christmas and New Year, while the hole in *September* is caused by 9/11.

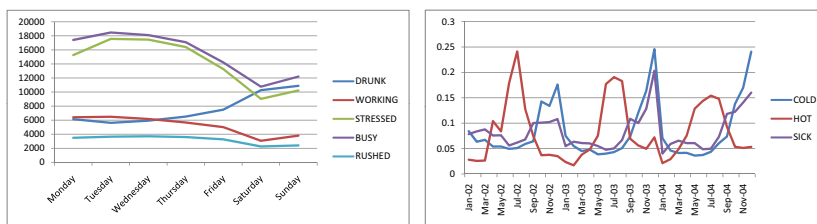


Fig. 4: The set of the lowest entropy moods. Left: The numbers of being tagged by weekdays; Right: The monthly proportion of being tagged.

³ <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/>

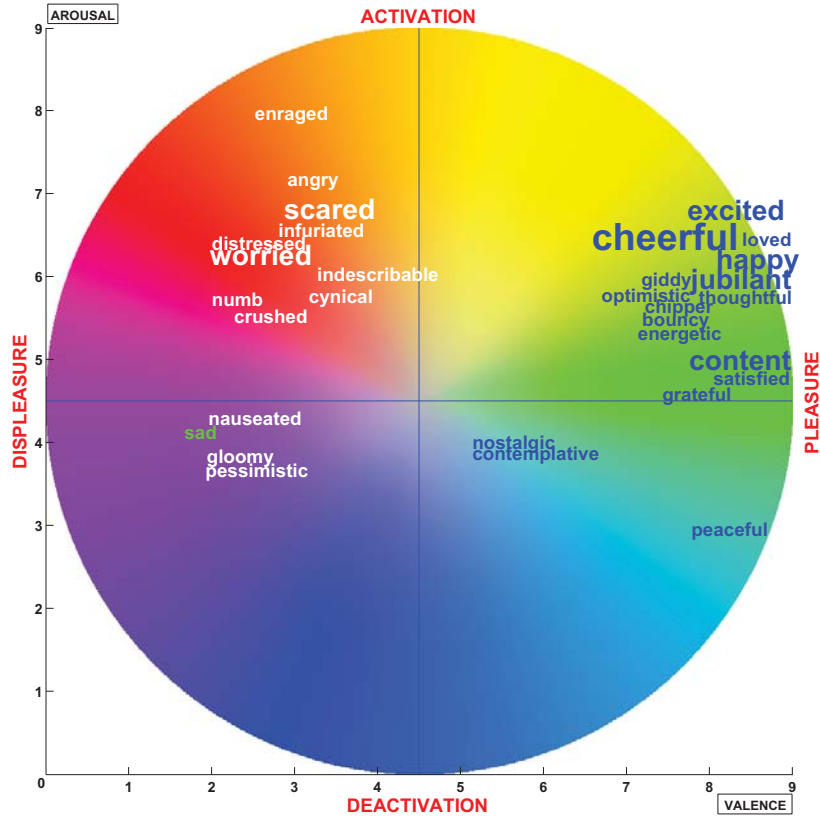


Fig. 5: Illustration of extracted moods on the affect circle (adapted from [10,11], added color-coding). Those in *blue* are for periodic events, in *white* for non-periodic events, and in *green* ('sad') for both.

2.2 Event indicators

In Section 2.1, we have considered sentiment in the blogosphere as a whole, aggregating over all moods, to detect sentiment-based events at a global level. However, each individual mood might have its own implications for detecting events. For example, a big drift in the mood *'shocked'* might indicate a catastrophic event, unlike a drift in the mood *'curious'*. Also, some moods might have cyclic emerging patterns over time, e.g. *'thankful'* on Thanksgiving Day or *'loved'* on Valentine's Day, while others may be increasingly tagged when non-periodic events happen, e.g., *'shocked'* for the 9/11 event. Motivated by this, we wish to detect which moods can potentially be used as good indicators for periodic and aperiodic events, in a data-driven manner.

| Day | Mood | Event |
|--------------|---|-----------------|
| 15 February | loved | Valentine's Day |
| 12 September | sad | 9/11 event |
| 24 December | excited | Christmas |
| 25 December | cheerful, chipper, giddy, grateful, happy, jubilant, peaceful | Christmas |
| 26 December | content, satisfied | Christmas |
| 31 December | bouncy, contemplative, energetic, nostalgic, optimistic, thoughtful | New Year |

Table 1: The set of moods found to reach their peaks during the time of periodic events.

| Day | Mood | Event |
|-------------|---|------------------|
| 11 Sep 2001 | pessimistic, scared | The 9/11 attacks |
| 12 Sep 2001 | angry, crushed, distressed, enraged, indescribable, infuriated, nauseated, numb, sad, worried | The 9/11 attacks |
| 11 Sep 2002 | cynical, gloomy, indescribable, sad | 9/11 anniversary |
| 20 Mar 2003 | scared, worried | Iraq war |
| 03 Nov 2004 | angry, crushed, cynical, distressed, enraged, gloomy, infuriated, nauseated, numb, pessimistic, scared, worried | 2004 US election |

Table 2: The set of moods found to reach their peaks during the time of aperiodic events.

Indicative moods for periodic events or habits We conjecture that a periodic event or bloggers' habits can cause a cyclic rise of tagging certain moods. Thus, those moods having low entropy – computed on the distribution of times they are tagged by a cycle, e.g. weekdays or months – potentially indicate periodic event patterns.

On the weekday case, we denote by $x^m = \{x_1^m, \dots, x_7^m\}$ the numbers of blog posts tagged with the mood m on $\{Monday, \dots, Sunday\}$ over the period. This vector is normalized so that $\sum_{i=1}^7 x_i^m = 1$. Then, the entropy for this proportion is computed by

$$\mathcal{H}(m) = - \sum_{i=1}^7 x_i^m \log_2(x_i^m)$$

The data used in this analysis ranges from 01 May 2001 to 27 December 2004 (191 weeks). The five lowest entropy moods detected are shown in Figure 4. '*Busy*' and '*stressed*' are quite similar in tagging distribution, peaking on Tuesday and gradually decreasing for the following days ('*Tuesday at 11:45 is most stressful time of the week*'⁴). '*Working*' and '*rushed*' are mostly used at the beginning of the week. '*Drunk*' is far different with the others since it is at the holes during early of the week, reaching the peaks at weekends.

In the same way, we detect for which months a given mood is increasingly tagged and infer periodical events with respect to the months. The data for this experiment

⁴ <http://www.telegraph.co.uk/news/newstopping/howaboutthat/5113653/Tuesday-at-1145-is-most-stressful-time-of-the-week-survey-suggests.html>

| Mood | Bursty period | Top topic |
|-------------|---------------|--|
| P*ssed off | 11Sep – 13Sep | world trade died center war country terrorists lost united states lives american die innocent attack government buildings live middle pentagon |
| Shocked | 11Sep – 15Sep | world country families family lost lives america shock tragedy trade innocent pray children events prayers die victims horrible scared ones |
| Angry | 11Sep – 16Sep | world plane planes center trade crashed towers pentagon tower twin scared tuesday low canton hour wtc hijacked tv building hit |
| Sympathetic | 11Sep – 19Sep | world lost pray families lives goes peace country family wish terrible died stop victims stand helping terrorism horrible loss tragedy |
| Worried | 11Sep – 21Sep | war world big america city end lost country lives president pray nation ones bush scared heard died scary planes fact |
| Enraged | 11Sep – 25Sep | lost heart family nation center wtc pentagon planes buildings war prayers act lives thoughts trade lucky reason race crashed hell |

Table 3: The set of moods found bursty in the 9/11 event and the related topics found in the blog posts tagged with the bursty moods in the bursty periods.

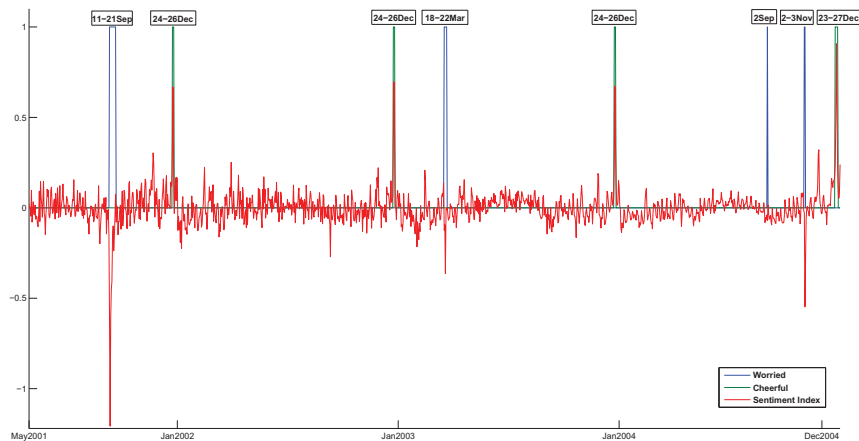


Fig. 7: Examples of moods found bursty throughout the time of the peaks and the holes of the sentiment index.

The essential idea of KLB is to model bursting as a generative process using a finite state-machine. Denote by s the state of the automaton, in the simple two-state model, one state is responsible for generating blog posts during non-burst period ($s = 0$) and another state is when the burst occurs ($s = 1$). Assume that there are n time points in our corpus, r_t is the number of blog posts tagged with a given mood at time t out of a total of d_t . The emission probability of the pair $\{r_t, d_t\}$ is modeled as a Binomial distribution together with a state-transition cost. A state sequence $\mathbf{q} = (q_1, \dots, q_n)$ of minimum total cost can be computed by dynamic programming. We refer readers to [8] for details of the algorithm.

A mood is considered bursty over a period if in the interval the corresponding state sequence is in high state. From the bursty intervals of given moods, we study if any bursty events are associated. We retrieve the blog posts tagged with the bursty moods in the bursty time to learn the events correlated. We use LDA for learning topics discussed in the content, as in Section 2.

3.2 Experimental results

We apply a two-state KLB automaton on the corpus to detect which moods are bursty and their bursty periods. 76 moods are detected bursty, resulting in 298 bursty intervals. ‘*Drunk*’, ‘*cold*’, and ‘*hot*’, as shown in Figure 6, have many bursty times but they are not found bursty in the time of the events detected in Section 2.

Other moods are found bursty in the time of the abnormal events. For example, six moods are found bursty on the 9/11 attacks and all top topics returned by LDA on the posts tagged with these moods during their bursty time mention the event (Table 3). Consistent with the finding in Section 2, a majority of these moods are low in *valence* and high in *arousal*.

Those moods found to have peaks during the periodic events in Section 2 are also detected bursty through these days. For example, ‘*cheerful*’, which has peaks on the 25th December in four years of the corpus, is also found bursty four times during Christmas (Figure 7). The top topics in the blog posts tagged with this mood in these bursty time are closely related to Christmas.

The correlation between the extreme points of the sentiment index and the bursty time of some moods (examples shown in Figure 7) complements the process of detecting events. While the former can help find the start time of events, the latter can help determine the events’ intervals.

Based on the proposed method, we implemented a system for querying events associated with bursts during specific time-periods using moods as keywords. Figure 8 displays screen-shots from the Web user interface (Web UI). A user can enter a set of moods (8a) and the software subsequently returns bursts associated with those moods along time-line for the specified duration. For example, a query of ‘*angry, sad*’ and duration ‘11th Sept 2001 to 31st Dec 2004’ results in 3 and 8 bursty events corresponding to the moods ‘*angry*’ and ‘*sad*’ respectively. Figure 8b provides a detailed result including topics and entities for the event ‘*9/11 attacks*’ and the time-line for two other events associated with the mood ‘*angry*’, including ‘*Iraq war*’ and ‘*US election*’.

⁵ <http://edition.cnn.com/2005/HEALTH/11/14/cold.chill/index.html>

Sentiment Burst Detection and Retrieval

Mood

angry,sad Browse Mood

Start Date / / **End Date** / /

(a) The input Web UI.

Burst Query Result

Record Expand All Collapse All

Topics Probability

Topics per Burst ✓ Max: 10

Links per Topics ✓ Max: 10

o Mood **angry**: 3 burst(s) detected

| Burst Id | Start | End | Length | Weight |
|----------|-------------|-------------|--------|--------|
| 11 | 11-Sep-2001 | 16-Sep-2001 | 6 | 80 |
| 12 | 18-Mar-2003 | 20-Mar-2003 | 3 | 17 |
| 13 | 03-Nov-2004 | 05-Nov-2004 | 3 | 526 |

o Mood **sad**: 8 burst(s) detected

world plane planes center trade **Location**

crashed towers pentagon tower ...

URG <http://www.kijournal.com/users/jessieq16948.html> Cached

The world is boiling... it will explode <http://www.kijournal.com/users/emmyko/2723.html> Cached

america country world americans terrorists attack ... **Organisation**

"These are the Times that try men's souls" - Thomas Paine <http://www.kijournal.com/users/coop316/10747.html> Cached

Terrorism <http://www.kijournal.com/users/cule/25219.html> Cached

war hell kill bomb countries Ming hurt give saved ... **People**

Fucking Scary Ass Site <http://www.kijournal.com/users/dreamtboys/7308.html> Cached

What makes people do shit like that? <http://www.kijournal.com/users/kat1234/1294.html> Cached

America, america, New_York, American, United_States, Washington, US, american, Canton, New_York_City, Pittsburgh, new_york, Iraq, Middle_East, Pearl_Harbor, Afghanistan, Akron_Canton, Alida, America_Wemad, Ansett,

Pentagon, WTC, Ioveable, CNN, ABC, American_Dream, United_Healthcare, AA, CCN, Catholic, Dirty_Abic_Records, ESPN, III, LOL, LOL_Ioveable, MTV, NYC, USA, USAir, WHO,

Liz, Steph, God, Bush, Alida, Jim, Larry, Mel, Arab, Aragon, Ben, Bin_Laden, Carol, David, Don, Mr_Cherry, Natalie, Sandra_Cisneros, Tom_Clancy, bush,

(b) The output Web UI.

Fig. 8: A system for querying bursty moods and related events.

4 Conclusion

We have investigated the novel problem of detecting sentiment-based bursty events in the blogosphere using mood tags. We proposed an approach to compute a sentiment index which can be used to extract subtle events in the stream of blogposts. We also found some moods themselves can be used as strong indicators of real-world events. While the sentiment indices can help find the start time of events, the bursty periods of moods returned by a state-machine can help to infer the bursty intervals of related events. The results have shown that the emotion information tagged can be exploited to detect meaningful events in blogs. We have also implemented a prototype Web interface for querying sentiment-based bursty events. Our proposed approach is generic and can be applicable to other streaming data.

References

1. D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
2. M.M. Bradley and P.J. Lang. Affective norms for English words (ANEW): Instruction manual and affective ratings. *University of Florida*, 1999.
3. R. Coontz. Blogs: Happiness barometers? *Science*, 325:5941, 2009.
4. J.R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proc. of the Association for Computational Linguistics Conference*, page 370, 2005.
5. E. Gilbert and K. Karahalios. Widespread worry and the stock market. In *Procs. of the Int. AAAI Conf. on Weblogs and Social Media (ICWSM)*, 2010.
6. J. Giles. Blogs and tweets could predict the future. *The New Scientist*, 206(2765):20–21, 2010.
7. T.L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(90001):5228–5235, 2004.
8. J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.
9. B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.
10. J.A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980.
11. J.A. Russell. Emotion, core affect, and psychological construction. *Cognition & Emotion*, 23(7):1259–1283, 2009.
12. A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Procs. of the Int. AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.