Decomposing Bloggers' Moods

Towards a Time Series Analysis of Moods in the Blogosphere

Krisztian Balog and Maarten de Rijke
ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
kbalog,mdr@science.uva.nl

ABSTRACT

Using a total of 20 million mood-annotated blog posts harvested between June 2005 and March 2006, we provide a time series analysis of the number of blog posts annotated with a mood. State-space methods are used to determine decompositions of the time series data associated with bloggers' moods (either individual or aggegrated), allowing us to look for patterns of trend, seasonality and cycle.

Our analysis reveals a broad spectrum of phenomena: (i) there is a clear overall decline in the usage of mood annotations; (ii) weather phenomena and holidays have a clear impact on the profile of some moods; (iii) looking at the relative counts, we observe that some moods are stationary, while others decline or climb; and (iv) several moods display changes in their cyclical or seasonal component during the period covered by our data.

Categories and Subject Descriptors

H.3.4 [**Information Systems**]: Systems and Software

General Terms

Blogs, moods, time series analysis

1. INTRODUCTION

Blogs, diary-like web pages containing highly opinionated personal commentary, are becoming increasingly popular. Many blog authoring environments allow bloggers to tag their entries with highly individual (and personal) features, offering us a unique look into people's reactions and feelings towards personal experiences and current events. Specifically, users of LiveJournal, one of the largest weblog communities, have the option of reporting their *mood* at the time of the posting; users can either select a mood from a predefined list of common moods such as "amused" or "angry," or enter free-text.

A large percentage of LiveJournal users choose to tag their posts with a mood. This results in a stream of weblog posts tagged with mood information per minute, from hundreds of thousands of users across the globe. We have been monitoring this stream and recording the number of blog posts annotated with a given mood at ten minute intervals since the summer of 2005, thus obtaining an ordered sequence of values at equally spaced time intervals for every mood. Our aim in this paper is to take a first step towards analyzing these time series. Time series analysis [3] deals with the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for. The usage of time series models is twofold: (i) Obtain an understanding of the underlying forces and structure that produced the observed data; (ii) Fit a model and proceed to forecasting, monitoring or even feedback and feedforward control.

In previous work we have reported on machine learning experiments aimed at the latter goal, forecasting aggregate mood levels [14]. Our interest in this paper lies with the first usage of time series analysis. Specifically, we want to determine general time series characteristics for all moods, both aggegrated and at the level of individual moods, to look for patterns of trend, seasonality and cycle.

Our analysis reveals a broad spectrum of phenomena: (i) there is a clear overall decline in the usage of mood annotations; (ii) weather phenomena and holidays have a clear impact on the profile of some moods but leave others unaffected; (iii) looking at the relative counts, we observe that some moods are stationary, while others decline, and still others climb; and (iv) several moods display changes in their cyclical or seasonal component during the period covered by our data.

The rest of the paper is organized as follows. In Section 2 we discuss related work. In Section 3 we describe the data used in this paper, as well as the time series analysis approach adopted. Then, in Section 4 we report on our findings; we present decompositions of many moods, and present a first attempt at grouping bloggers' moods by their time series characteristics. We conclude in Section 5.

2. RELATED WORK

Mood related work can be viewed as a special type of sentiment analysis. Classifying the mood of a single blog post is a hard task; state-of-the-art methods in text classification achieve only modest performance in this domain [13]. As to tracking moods or sentiments in blogs across extended periods of time, the activity and trend watching services that (blog) search engines such as BlogPulse provide [8] are related to our work, but different. Recently, a time series perspective was adopted for targeted online tracking of chatter (in blogs and elsewhere) around a specific set of topics or products [9]. Mishne and Glance [16] test how well sentiment in blogs performs as predictor for future sales information.

Mood forecasting (using linear regression techniques) has been studied in [14], leading to implementations described in [15]. Correlations between textual features and a very

WWW2006, May 22–26, 2006, Edinburgh, UK.

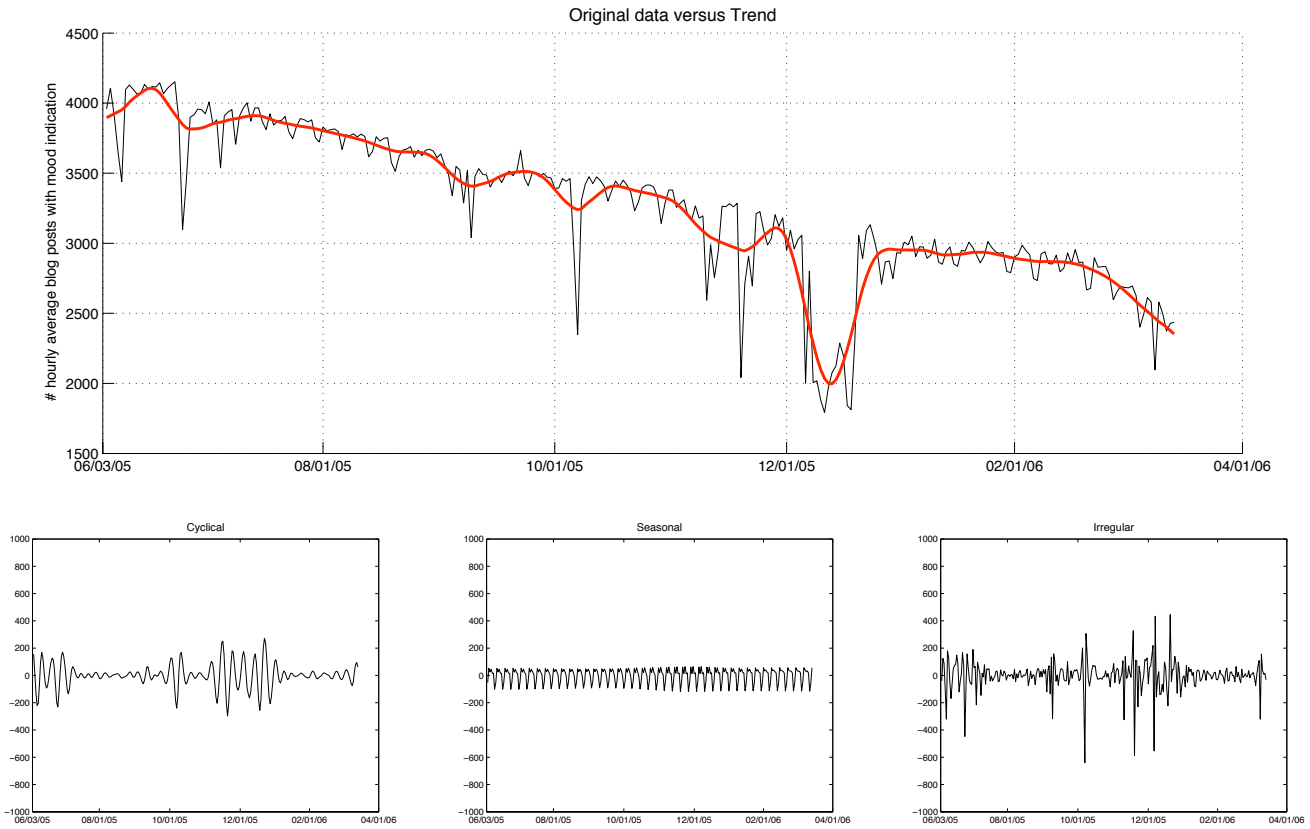

Figure 1: Decomposition of the average number of hourly blog posts with mood annotation. (Top): Original data vs the trend; (Bottom-Left): Cyclical component; (Bottom-Center): Seasonal component; (Bottom-Right): Irregular component. (Unlike later plots for individual moods, for these overall counts, we used the number of hourly posts).

specific mood (*happiness*) are examined in [12], while more general correlations between unusual fluctuations in mood levels and global news events are studied by [1].

Finally, Mei *et al.* [11] present general probabilistic methods for discovering and summarizing the evolutionary patterns of themes in blogs by discovering latent themes, constructing an evolution graph of themes, and analyzing life cycles of themes.

3. EXPERIMENTAL SETUP

We describe our data as well as the time series analysis methods employed.

3.1 Data

Our analysis is based on a collection of blog posts from LiveJournal.com, annotated with moods. LiveJournal.com blogs are generally considered to be "personal" blogs, closely reflecting the daily events and experiences of LiveJournal bloggers. As the indication of the mood is optional when posting on LiveJournal, the mood-annotated blog posts we use are likely to reflect the true mood of the bloggers.

Our corpus consists of all public blogs published in LiveJournal during a period of approximately 280 days, from June 2005 to March 2006. The moods used by LiveJournal users are either selected from a predefined list of 132 moods, or entered in free-text. Posts without an explicit mood indication as well as posts with free-text mood indications are discarded, leaving us with a total of 20 million blog posts.

Figure 1 (Top) shows the average number of hourly mood-annotated blog posts. Somewhat surprisingly, the figure illustrates that the number of blog posts that have a mood annotated, displays a decreasing trend, and this trend emerges in the individual mood levels as well. This fact motivated us to use relative mood counts instead of absolute ones.

For each mood, (relative) time series data was generated based on the following formula (applied for each individual day):

$$val(mood, day) = \frac{day\_avg(mood)}{\sum_{m \in moods} day\_avg(m)}, \quad (1)$$

where $moods$ is a set of pre-defined moods and $day\_avg(m)$ denotes the average daily number of blog posts annotated with mood m.

3.2 Decomposition Method

We first present our general modeling decisions, and then detail our choice of parameters and parameter estimation.

We use an additive structural decomposition of a time series z_t in terms of four components:

$$z_t = t_t + s_t + c_t + i_t, \quad (2)$$

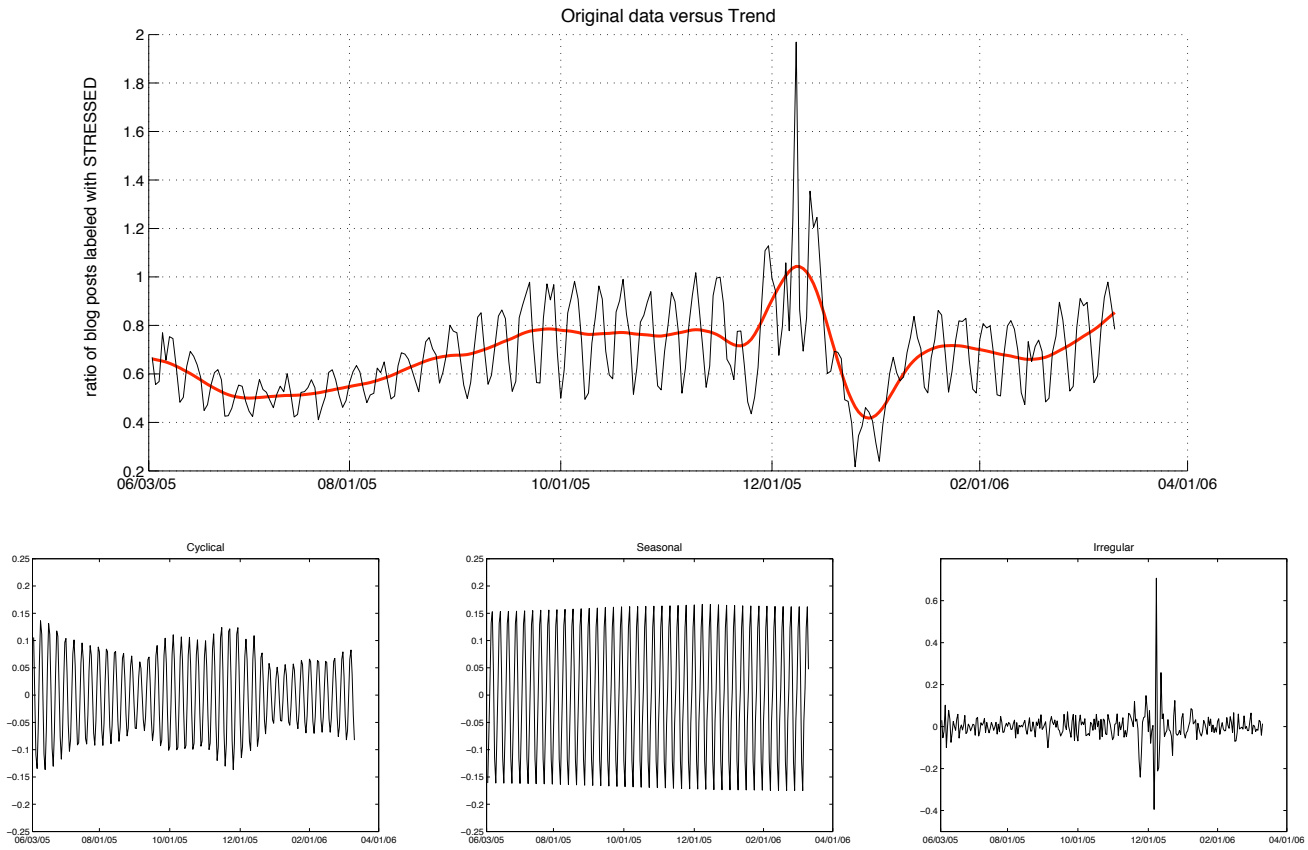

Figure 2: Decomposition of the mood *stressed*. (Top): Original data vs the trend (Bottom-Left): Cyclical component, (Bottom-Center): Seasonal component, (Bottom-Right): Irregular component. The values represent the ratio of blog posts labeled with the mood *stressed*.

where

- t_t is the *trend component*, representing the long-term behavior of the series,
- s_t is the *seasonal component* which is dependent on the time of the year and describes any regular fluctuations with a period of less than one year,
- c_t is the *cyclical component*, describing regular fluctuations apart from seasonal effects, and
- i_t is the *irregular component*.

There are two basic approaches to estimating the components on the right hand side of (2): *ad-hoc* and *model-based*. *Ad-hoc* methods filter the series by means of a differential equation. A clear advantage in usability is that these methods do not require any previous analysis. However, these procedures implicitly assume that different time series follow the same stochastic process. A major problem is that the estimates change when the sample increases ("revision") [17].

Model-based methods were inspired by the potential inadequacy of ad-hoc methods and emphasize the coherence between the properties of the observed series and those of the structural components. Several methodologies are known, here we restrict ourselves to the *ARIMA model-based* and the *Structural Time Series Models* ("structural models") approaches. ARIMA-model-based techniques [2, 10] first build an ARIMA representation of z_t, and then obtain a structural decomposition defined by individual ARIMA processes for each component, constrained by the requirement that their sum is equivalent to the model for the time series [7]. Structural models share many elements with the ARIMA approach, as they rely on similar hypotheses and use equivalent extraction techniques. The most important feature of structural models, in comparison with common practice and most alternative procedures, is that these components are not assumed to be orthogonal conditional on their past. Components are represented by space-state models specified according to the properties of the time series. The space-state representation assures that the estimates of the components converge to values with null variances and covariances. Null variances guarantee that the components do not change when the size of the sample increases. On the other hand, null covariances assure that a given component can be analyzed and interpreted independently from any other component(s).[1]

For decomposing the time series associated with moods we arrive at a structural model by adding several dynamic struc-

[1]For a more complete account of structural time series models and state-space representation we refer the reader to [5].

tures, each of which has a simple underlying model that describes its particular dynamic and stochastic features. Trend, cyclical and seasonal components were modeled by ARIMA processes while the irregular component was modeled by white noise. The general ARIMA model introduced by Box and Jenkins [2] may be summarized as ARIMA(p, d, q) where the three types of parameters in the model are: the autoregressive (AR) parameters (p), the number of differencing passes (d) and moving average (MA) parameters (q). The seasonal part of an ARIMA model has the same structure as the non-seasonal part: it may have an AR factor, an MA factor, and/or an order of differencing. In the seasonal part of the model, all of these factors operate across multiples of lags (the number of periods in a season). A seasonal ARIMA model is classified as an ARIMA(p,d,q)x(P,D,Q) model, where P is the number of seasonal autoregressive (SAR) terms, D is the number of seasonal differences, and Q is number of seasonal moving average (SMA) terms.

First, we need to decide on the specific number and type of ARIMA parameters to be estimated. The decision is not straightforward, since ARIMA is a complex technique and requires a great deal of experience. An ARIMA(2,2,0) model was fitted for trend and cyclical components, while the seasonal component was modeled by ARIMA(0,1,1)x(0,1,1) and the length of seasonality was set to 7.

Once the specific number and type of parameters has been decided upon, the next step is to estimate their values. During the parameter estimation phase a function minimization algorithm is used to maximize the likelihood of the observed series, given the parameter values. The numerical optimization procedure we use is based on the Broyden-Fletcher-Goldfarb-Shanno algorithm [4].

Our experiments were conducted using E^4, a MATLAB toolbox for time series modeling and decomposition [6]. E^4 estimates econometric models by exact maximum likelihood. It supports space-state representation and structural time series models.

4. FINDINGS

Our data is very rich: we track a total of 132 moods over a period of approximately 280 days. Below, we present a selection of our time series analysis, organized in five groups: global analysis, zooming in on a particular mood (*stressed*), the impact of global events, different types of trend, and, finally, moods with changes in their cyclical or seasonal component during the period covered by our data.

4.1 Global analysis

Returning to Figure 1, the trend (the red curve in the top figure) is clear: it is downward. The number of blog posts with a mood annotated has been dropping since the start of the period covered by our data. The cyclical component in Figure 1 shows increased oscillations around the time of the London bombings in early July 2005—in the irregular component we observe increased oscillations around the same periods. The seasonal component is uneventful.

4.2 Zooming in on a single mood

By way of illustration, we zoom in one particular, *stressed*. Figure 2 (Top) shows the decomposition of this mood. The original data shows increasingly steep (weekly) peaks as we move into the winter and the end of year period, with stress levels peaking at the very end of the year. The "interesting" phenomenain the raw and irregular occur at the same times, while the cyclical component shows a decline moving towards the fall, and then a rise moving towards the end of the year (peaking at the same point in time as the raw and irregular components), which is then followed by a drop at the start of the year. The cyclical plot displayed in Figure 2 (Bottom-Left) shows more intense oscillations then, e.g., the cyclical component graph for *cold* (Figure 5 (Top-Right)). The seasonality graph for *stressed* is very constant. The peaks in the *Irregular* component co-occur with end of year deadlines.

4.3 The impact of global events

We found that global events and holidays have a profound influence on mood levels. As an example, Christmas and New Year's Day impacted almost all moods; see Figure 3, where we display the original data and trend for *cheerful* and *annoyed*.

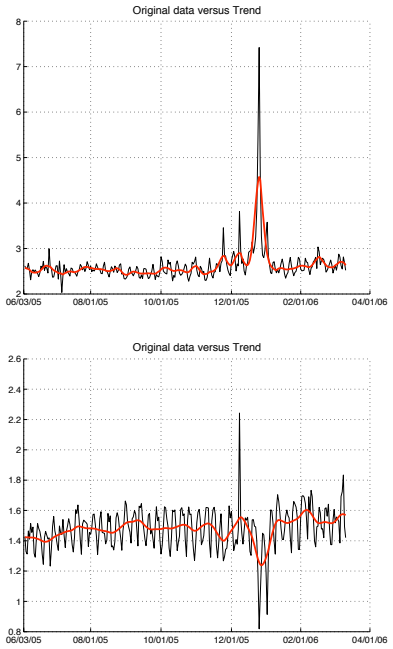

Figure 3: The impact of global events. (Top): cheerful, (Bottom): annoyed

4.4 Four types of trend

We examined the trend types for all moods and identified four types of trend: irregular (Figure 2, illustrated with *stressed*), constant (Figure 4 Left, illustrated with *drunk*), constantly decreasing (Figure 4 Center, illustrated with *excited*) and constantly increasing (Figure 4 Right, illustrated with *contemplative*).

4.5 Changing cyclical or seasonal components

Finally, we consider examples of moods that display changes in their seasonal and/or cyclical components. Before we zoom into concrete examples, we briefly review the difference between these two components. The seasonal component describes variation that is annual in period. Apart

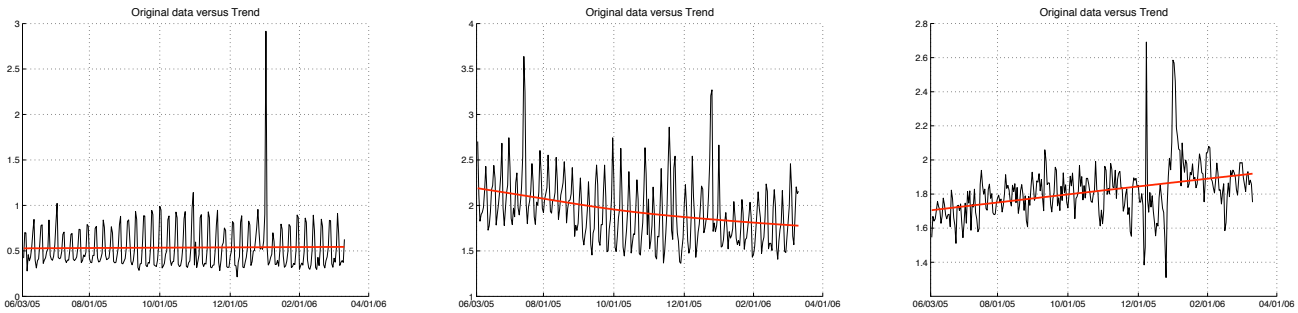

Figure 4: Different trend types. (Left): constant for *drunk*, (Center): monotonically decreasing for *excited*, (Right): monotonically increasing for *contemplative*

from seasonal effects, time series may exhibit variation at a fixed period. The cyclical component represents these 'other' cyclic variations. In addition, the cyclical component may describe fluctuations, which do not have a fixed period, but are predictable to some extent. Since we have less than a year's worth of data, we are not able yet to capture true 'seasonal' variations.

Looking at *cold*, we see changes in both components around periods of increased cold. The mood *sleepy* provides a nice example of a mood with a changing cyclical component: it changes dramatically at the start of the school year, around the start of the school year and/or academic year in late August or early September; see Figure 5.

5. CONCLUSIONS

Using a total of 20 milion mood-annotated blog posts harvested between June 2005 and March 2006, we provide a time series analysis of the number of blog posts annotated with a mood. We found that, overall, there is a downward trend in the number of posts with mood annotation. Moreover, we found a broad spectrum of phenomena: (i) weather phenomena and holidays have a clear impact on the profile of some moods but leave others unaffected; (ii) looking at the relative counts, we observe that some moods are stationary, while others decline, and still others climb; and (iii) several moods display changes in their cyclical or seasonal component during the period covered by our data.

As to future work, as this paper describes early research into a time series perspective on bloggers' moods, we believe that there is potential mileage in more complex modelings of the data, ones that take into account (possible) daily, weekly and monthly intra- and interdependencies. While we have used standard models that seem to fit our data well, there is room for improvement.

Acknowledgements

This research was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 612-13-001, 612.000.106, 612.000.207, 612.066.302, 612.069.006, 640.001.501, and 640.002.501.

6. REFERENCES

[1] K. Balog, G. Mishne, and M. de Rijke. Why are they excited? Identifying and explaining spikes in blog mood levels. In *Proceedings 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, 2006.

[2] G. E. P. Box, S. C. Hillmer, and G. C. Tiao. Analysis and Modeling of Seasonal Time Series. In *Seasonal Analysis of Time Series, ed. A. Zellner, Washington, DC: Bureau of Census*, 1978.

[3] P. Brockwell and R. Davis. *Introduction to Time Series and Forecasting*. Springer, second edition, 2003.

[4] J. E. Dennis and R. B. Schnabel. Numerical methods for unconstrained optimization and nonlinear equations. In *Englewood Cliffs (N.J.): Prentice-Hall*, 1983.

[5] J. Durbin and S. Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2001.

[6] E^4. State-Space Estimation of Econometric Models, 2001. http://www.ucm.es/info/icae/e4/.

[7] V. Gómez and A. Maravall. Programs TRAMO and SEATS: Instructions for the User. In *Working Paper 9628, Madrid: Bank of Spain*, 1996.

[8] N. Glance, M. Hurst, and T. Tomokiyo. Blogpulse: Automated trend discovery for weblogs. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004*, 2004.

[9] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 78–87, 2005.

[10] S. C. Hillmer and G. C. Tiao. An ARIMA-Model-Based Approach to Seasonal Adjustment. *Journal of the American Statistical Association*, 77:63–70, 1982.

[11] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the World Wide Web Conference 2006 (WWW'06)*, To appear.

[12] R. Mihalcea and H. Liu. A corpus-based approach to finding happiness. In *AAAI 2006 Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[13] G. Mishne. Experiments with mood classification in blog posts. In *Style2005 – 1st Workshop on Stylistic Analysis of Text for Information Access, at SIGIR 2005*, 2005.

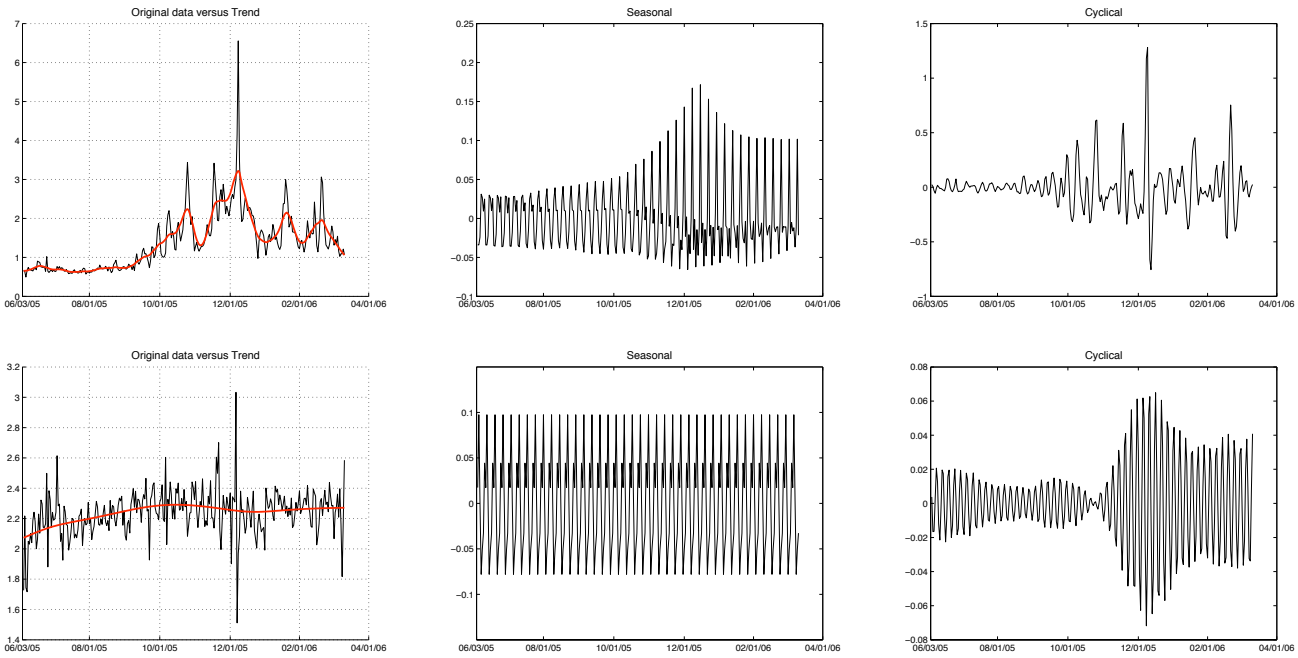

Figure 5: (Top): Decomposition of the mood *cold*, (Bottom): Decomposition of the mood *sleepy*. The components are (Left): Original data vs the trend, (Middle): Seasonal, (Right): Cyclical

[14] G. Mishne and M. de Rijke. Capturing global mood levels using blog posts. In *AAAI 2006 Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[15] G. Mishne and M. de Rijke. MoodViews: Tools for blog mood analysis. In *AAAI 2006 Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[16] G. Mishne and N. Glance. Predicting movie sales from blogger sentiment. In *AAAI 2006 Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[17] J. Shiskin, A. H. Young, and J. Musgrave. The X-11 Variant of the Census Method II Seasonal Adjustment Program. In *Technical Paper, Washington, DC: Bureau of the Census*, 1967.