Applied Text Analytics for Blogs

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof.dr. J.W. Zwemmer
ten overstaan van een door het college voor
promoties ingestelde commissie, in het openbaar
te verdedigen in de Aula der Universiteit
op vrijdag 27 april 2007, te 10.00 uur

door

Gilad Avraham Mishne

geboren te Haifa, Israël.

Promotor: Prof.dr. Maarten de Rijke

Committee:
Prof.dr. Ricardo Baeza-Yates
Dr. Natalie Glance
Prof.dr. Simon Jones
Dr. Maarten Marx

Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Universiteit van Amsterdam

SIKS Dissertation Series No. 2007-06
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

The investigations were supported by the Netherlands Organization for Scientific Research (NWO) under project number 220-80-001.

Printed and bound by PrintPartners Ipskamp, Enschede

ISBN-13: 978–90–5776–163–8

Chapter 6

Aggregate Sentiment in the Blogspace

Sentiment analysis is a complex task; typical performance in this domain is lower than that achieved in other, more straightforward text classification tasks, such as topical categorization. We have already observed, in Chapter 4, that classifying the mood of a blog post is hard; more subtle expressions of sentiment—e.g., irony or sarcasm—remain a challenge for current technology (as well as for many humans). But, as with many other computational linguistics tasks, overall performance of sentiment analysis techniques increases as more data is available: more training examples contribute to a better model, and longer texts in the training or test sets contribute to stability and accuracy of the method.

In this Chapter we turn to mining the personal, sentiment-rich language of blogs—this time at an aggregate level. Instead of analyzing a single post or blog, we look at the sentiment as reflected in multiple blogs, or the entire blogspace. We begin by demonstrating why this is useful: Section 6.1 examines the relation between aggregate blogger sentiment and financial results of products, showing that taking the sentiment into account results in better models than those obtained when measuring only the volume of discussion related to a product. The rest of this Chapter continues to explore moods in the blogspace—an area we addressed at the level of single posts in Chapter 4. Section 6.2 demonstrates that important insights can be obtained by observing moods reported by bloggers over many blogs. In Section 6.3 we develop a method for predicting the global mood through the language used by multiple bloggers, and in Section 6.4 we focus on irregular temporal patterns of such moods, and how they can be explained—again using the bloggers' language. As the latter part of the chapter will show, analysis of mood at the aggregate level is more tractable than the corresponding analysis at the level of a single post.

6.1 Blog Sentiment for Business Intelligence

Earlier, we referred to a blog as the "unedited voice of an individual." The entire blogspace, then, can be viewed as the voice of the public: a massive collection of discussions and commentary reflecting people's opinion and thoughts. Part of this discussion is classified as Consumer Generated Media (CGM, [34])—experiences and recommendations expressed about products, brands, companies and services. CGM in blogs presents a double opportunity: for consumers, this type of advice provides direct, unsolicited information that is often preferred to more traditional channels; a survey of online consumers showed that people are 50% more likely to be influenced by word-of-mouth recommendations than by radio or television advertisements [126]. For companies, CGM helps to understand and respond to the consumer by analyzing this informal feedback. This Section focuses on the latter use of CGM.

A relation between the volume of discussion in blogs (the "buzz") and commercial performance of a product has already been observed (e.g., [97]). In addition, sentiment analysis methods for analyzing typical CGM content have improved substantially in recent years, based on the availability of large-scale training data and resources. The main question addressed in this section is whether these two aspects can be combined: more specifically, we aim to discover whether usage of sentiment analysis on blog text in a CGM context results in a better predictor of commercial performance than simple buzz count. To this end, we analyze the sentiment expressed in blogs towards a particular product type—movies—both before the movie's release and after, and test whether this sentiment correlates with the movie's box office information better than a simple count of the number of references to the movie in blogs does. We proceed by describing related work, the data we used in our experiments, and their results.

6.1.1 Related Work

Product reviews are frequently used as the domain in sentiment analysis studies (e.g., [238, 61]); they are focused, easy to collect, and often provide meta-data which is used as ground truth: a predefined scale which summarizes the level and polarity of sentiment ("4 out of 5 stars"). Blogs differ from these studies in that they tend to be far less focused and organized than the typical product review data targeted by sentiment analyzers, and consist predominantly of informal text. Often, a reference to a movie in a blog does not come in the context of a full review, but as part of a post which focuses on other topics too.

A number of studies are closer to the work described here than most product review oriented sentiment work. Good correlation between movie success and blog posts mentioning the movie was established in [291]. However, this study was based on an analysis of five blogs only; furthermore, the tracked blogs are dedicated to movie reviews, and their content resembles professional product

review sites rather than typical blog content. It is unclear whether the methodology described scales up to other products and blogs, or whether it is different from simply tracking non-blog product reviews. A large-scale study of blogs and business data, measuring the correlation between the number of blog posts mentioning a product (books, in this case) and its sales rank, is described in [97]. Here, the raw number of product mentions in the blogspace was shown to be a good predictor of sales, but no sentiment analysis was used. Tong [292] and Tong and Snuffin [293] do describe systems which incorporate sentiment analysis for measuring correlation between business information and product mentions, but do not report on empirical results. Finally, in work done in parallel to that reported here, Liu analyzed references to movies in message boards, finding that sentiment is not particularly beneficial as an indicator of movie success [177]. In this study, most movie mentions were observed after the release of a movie; this correlates with our own observation regarding post-release discussions. However, we also analyze, separately, pre-release references to movies—reaching a different conclusion in this case.

6.1.2 Data and Experiments

Our task, then, is to examine the correlation between sentiment expressed in blogs towards a product, and the product's financial success; our product of choice is movies—as mentioned earlier, a popular domain for sentiment analysis studies. We begin by presenting the data and the methodology used for examining this correlation.

Product Set

To measure aggregate levels of sentiment, we first require the studied movie to have some minimal discussion volume in the blogspace: in cases where only a handful of references to a movie exist, there is little meaning to aggregating their infomation. For this reason, we limited our analysis to high-budget movies—requiring a budget higher than 1 million U.S. dollars. During the period between February and August 2005, which is the time span we study, 49 movies with publicly-available financial information meet this criterion; these are the movies used in the work that follows.

Financial Data

Given a particular movie, we used IMDB—the Internet Movie Database[1]—to obtain the date of its "opening weekend" (the first weekend in which the movie played in theaters), as well as the gross income during that weekend and the number of screens on which the movie premiered. We focus on the opening

[1] http://imdb.com

weekend data rather than total sales since this normalizes the figure across movies that were released on different dates, preventing earlier movies from having a higher total income just because they have been "out there" longer, have been already released on DVD, etc. Opening weekend income correlates highly with total movie income, accounting for an estimated 25% of the total sales [283]. The number of screens the movie premiered on was used to normalize the opening weekend income, producing an "Income per Screen" figure for each movie. This allows comparing sales of summer blockbuster movies, sometimes released to 4,000 screens simultaneously in the opening weekend, with lower-profile movies released to 1,000–2,000 screens.

Blog Data

For each movie, we collected all related blog posts appearing in the Blogpulse [90] index, a large-scale index covering the majority of the blogspace. A post was considered related to a movie if the following conditions were true:

- The date of the post is within a window starting a month prior to the movie's opening weekend date and ending one month after it.
- The post contained a link to the movie's IMDB page, *or* the exact movie name appeared in the post in conjunction with one of the words ⟨movie, watch, see, film⟩ (and their morphological derivatives).[2]

Methodology

For each relevant post, we extracted the contexts in which the movie was referenced by taking a window of k words around the hyperlinks to the movie's IMDB page, or around exact matches of the movie name; we used varying values for k, from 6 words to 250. Then, we calculated the sentiment polarity and intensity of the contexts using the state-of-the-art method described in [222]. In this approach, a domain-specific lexicon of sentiment values is used to tag each token in the text; we used a lexicon developed for the movie domain. Additionally, the text is chunked, and predefined expressions of polarity are located in the higher-level groups of tokens; finally, polarity information is propagated to the entire sentence level, using heuristics to handle negation, subsumption, and so on. This method has been shown to perform only a few percent worse than human performance on the same task [222].

Examples of different context lengths for the same reference to a movie are shown in Table 6.1, along with the computed polarity; this demonstrates the

[2]This strategy aims for high precision without overly sacrificing recall. An examination of the retrieved posts shows a high degree of precision. We did not explicitly measure recall, but did find that using a combination of an IMDB link query and text queries increased recall by a factor of 500% over simply using the IMDB link query, which has near-perfect precision but limited recall.

Length (k)	Content	Sentiment
8	Rach went to see "The Sisterhood of the Traveling Pants", we both read the	Neutral
40	gym or work. 2. It was the first time I laughed since I've been home. Me and Rach went to see "The Sisterhood of the Traveling Pants", we both read the book and the girl from Gilmore Girls was in it. I had the best time	Positive
120	Tonight was a night of firsts. 1. I'm pretty sure it was the first time I left the house and didn't go to the gym or work. 2. It was the first time I laughed since I've been home. Me and Rach went to see "The Sisterhood of the Traveling Pants", we both read the book and the girl from Gilmore Girls was in it. I had the best time with her. We laughed, were both such dorks. The movie was SOOO sad. Like true "The Notebook" quality. I enjoyed it and it got me thinking. I need to stop being so miserable. I make my time here suck. I	Negative

Table 6.1: Polarity of different contexts.

possible differences in polarity estimation when using "too much" or "too little" context.

In summary, for every item m in the set of 49 movies we have the following information:

- d_m: the opening weekend date of m;
- e_m: earnings from ticket sales for m during the opening weekend;
- s_m: the number of screens m featured on in the opening weekend;
- R: a collection of references to m in blog posts. For each $r \in R$, we also have d_r—the date of the post containing r; and $p_{k,r}$, the polarity value of the k words surrounding r, where k values vary between 6 and 250.

A sample item is shown in Table 6.2. Note that the polarity score is fitted to a log-linear distribution, with the majority of scores falling within a range of 4 to 7 [223]. Thus, the average polarity score of 5.5 for the movie in the table indicates significant positive overall sentiment.

Using the d_r values, we can partition R into two subsets: R_{pre} which is all references made to m prior to its release (i.e., $d_r < d_m$), and R_{post}—all references made after the release ($d_r \geq d_m$). Then, we can measure the correlation between $|R_{pre}|$ or $|R_{post}|$ and the "Income per Screen," e_m/s_m, as well as measure sentiment-related correlations that take into account the polarity values, p_r.

Movie	The Sisterhood of the Traveling Pants
Opening Weekend (d_m)	5 June 2005
Opening Weekend Sales (e_m)	\$9.8M
Opening Weekend Screens (s_m)	2583
Income per Screen (e_m/s_m)	\$3800
Pre-release Data	
References in blogs ($\|R_{pre}\|$)	1773
Context Length: 10 words	
- Positive references	329
- Negative references	50
- Mean sentiment polarity	5.5 / 10
Context Length: 20 words	
...	...
Post-release Data	
References in blogs $\|R_{post}\|$	1618
...	...

Table 6.2: Sample data from our collection.

Experiments

At this stage, we have some indicators of the movie's success (Income per Screen and raw sales figures), as well as a range of sentiment-derived metrics such as the number of positive contexts, the number of negative ones, or the total number of non-neutral contexts. The natural way to determine whether the two sets of information are related is to measure the statistical correlation between them; we use Pearson's r-correlation for this. In addition to measuring the statistical correlation between the sentiment-related measures and the movie success information, we measure the correlation between the raw counts of occurrences in blogs (the "buzz") and the financial information: comparing the two correlations will address the main question in this section: whether sentiment information improves on volume counts only for this type of task. Measurement was done separately for pre-release contexts and post-release ones.

Raw counts vs. sentiment values. Our first observation is that usage of the sentiment polarity values, given an optimal context length, results in better correlation levels with movie success than the raw counts themselves for data gathered *prior* to the movie's release. For data gathered *after* the movie's release, raw counts provided a better indicator. Of the different polarity-based measures used in our experiments, those yielding the best correlation values were as follows:

- Prior to the movie release: the number of positive references, within a relatively short context (the optimal value was 20 words).
- After the movie release: the number of non-neutral references within a relatively long context (the optimal length was 140). Using the number of

positive references achieved very close results to this.

Table 6.3 compares the correlation between movie business data for raw counts and for the best performing polarity-related metrics. Clearly, the sentiment-based correlation improves substantially over the raw counts for pre-release data, whereas for post-release data the effect is negative (but minor).

Correlation	**Between ...**	**Period**
0.454	Raw counts and income per screen	Pre-release
0.509 (+12%)	Positive contexts and income per screen	
0.484	Raw counts and sales	
0.542 (+12%)	Positive contexts and sales	
0.478	Raw counts and income per screen	Post-release
0.471 (-1%)	Non-neutral contexts and income per screen	
0.614	Raw counts and sales	
0.601 (-2%)	Non-neutral contexts and sales	

Table 6.3: Comparison of correlation between movie business data and blog references, with and without use of sentiment. Context sizes used: 20 (pre-release), 140 (post-release).

While the improvement using sentiment values on pre-release data is in-line with intuition, it is unclear to us why it does not have a similar effect for post-release data. One possible explanation is that post-release contexts are richer and more complex, decreasing the accuracy of the sentiment analysis.

Context length. Our next observation is that constraining the context being analyzed to a relatively small number of words around the movie "anchor" is beneficial to the analysis of pre-release polarity metrics, but reduces the effectiveness of the post-release metrics. Figure 6.1 displays the relation between the correlation values and the context length for two particular instances of analysis: the correlation between the number of positive contexts before the movie release and the income per screen, and the correlation between the number of non-neutral contexts after the release and the opening weekend sales for the movie (note that the context length is plotted on a log-scale).

Examining the contexts extracted both before and after the movie's release, we observed that references to movies before their release tend to be relatively short, as the blogger typically does not have a lot of information about the movie; usually, there is a statement of interest in watching (or skipping) the movie, and possibly a reaction to a movie trailer. References to movies after their release are more often accounts of the blogger's experience watching the movie, containing more detailed information—see an example in Table 6.4. We hypothesize that this may be an explanation for the different effect of context length on the correlation quality.

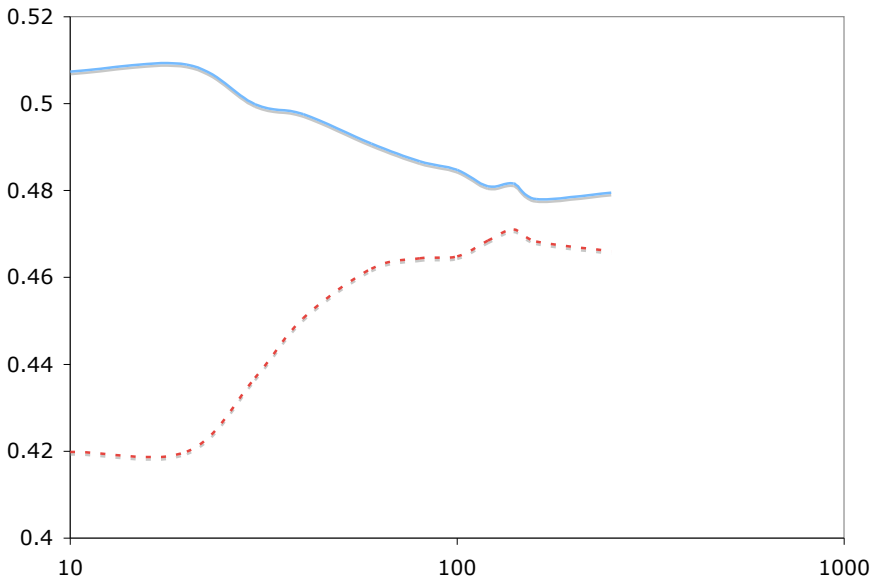

Figure 6.1: Relation between context length and correlation to income per screen: positive references, pre-release (blue, solid line) and non-neutral references, post-release (red, dashed line). The X-axis shows the context length (on a log-scale), and the Y-axis shows the level of correlation.

Breakdown

Out of the 49 movies in our study, over half have very good correlation between pre-release positive sentiment and sales. Less than 20% can be viewed as outliers: movies whose average Income per Screen was poorly predicted by pre-release sentiment. How can the low correlation between blog opinion and business data be explained for these outliers? Movie sales have been shown to be affected by many factors unrelated to online discussion, such as genre, Motion Picture Association of America rating, other movies released at the same time, and so on [283]. On top of that, noise originating from different components of our analysis—the retrieval of posts from the collection of all posts, the polarity analysis, and so on—accumulates, and may destabilize the data.

Cursory examination of outliers in our experiments, both those that overestimate sales and those that underestimate them, did not yield any obvious feature shared by the irregular data points.

6.1.3 Conclusions

The purpose of this Section was to motivate our work on aggregate sentiment analysis in blogs. To this end, the question we set out to investigate was whether taking into account the language used in blogs towards products—and, more

apparently an early easter is bad for apparel sales. who knew? i'll probably go see "guess who?" this weekend. i liked miss congeniality but the sequel [link to IMDB's page for "Miss Congeniality 2"] looks *awful*. and seattle's too much of a backwater to be showing D.E.B.S. i had to wait forever to see saved! too. mikalah gordon got kicked off american idol last night. while she wasn't the best singer, i wish ...
Monday, March 28, 2005 - Miss Congeniality 2: Armed and Fabulous. I know this is overdue, but I wanted to use this opportunity to discuss an important topic. The issue at hand is known as the Sandra Bullock Effect (SBE). This theorem was first proposed by my brother, Arthur, so he is the real expert, but I will attempt to explain it here. The SBE is the degree to which any movie becomes watchable simply by the presence of a particular actor or actress who you happen to be fond of. For example, if I told you that someone made a movie about a clumsy, socially awkward, dowdy female police officer who goes undercover as a beauty pageant contestant to foil some impenetrable criminal conspiracy, you'd probably think to yourself, "Wow that sounds pretty dumb." And you'd be right. However ...

Table 6.4: Typical references to movies in blogs: pre-release (top), and post-release (bottom).

concretely, using sentiment analysis to classify this language—results in a better prediction of the financial success of the products than measuring only the amount of discussion about them in the blogspace. The answer to this question is positive: in the domain of movies, there is good correlation between references to movies in blog posts—both before and after their release—and the movies' financial success; but shallow sentiment analysis methods can improve this correlation. Specifically, we found that the number of positive references to movies in the pre-release period correlates better with sales information than the raw counts in the same time period.

By itself, the correlation between pre-release sentiment and sales is not high enough to suggest building a predictive model for sales based on sentiment alone. However, our results show that sentiment might be effectively used in predictive models for sales in conjuction with additional factors such as movie genre and season. More generally, we conclude that aggregate sentiment analysis in the blogspace is useful for this type of tasks; blogs provide a unique source of CGM information through their personal, unmediated nature, and their sheer scale.

6.2 Tracking Moods through Blogs

We have already discussed moods in blogs in Section 4.1, which focused on identifying the mood of a given blog post. In the rest of this Chapter we return to blog moods, this time at the aggregate level. The mood assigned to a single post gives an indication of the author's state of mind at the time of writing; a collection of such mood indications by a large number of bloggers at a given point in time provides a "blogspace state-of-mind," a global view of the intensity of various feelings among people during that time. The "long tail" of the blogspace was described in Section 2.2.2: it is the body of blogs which comprises the majority of the blogspace. This section examines the the mood indications of the individuals which are part of this long tail, as expressed through their blogs over time.

Tracking and analyzing this global mood is useful for a number of applications. Marketers and public relation firms would benefit from measuring the public response to introductions of new products and services; political scientists and media analysts may be interested in the reaction towards policies and events; sociologists can quantify the emotional echo of an event throughout the crowds.

The rest of this Chapter offers methods for analyzing blog content to predict mood changes over time and reveal the reasons for them. But before applying text analytics to mine mood-related knowledge, we motivate our work by demonstrating the type of insights gained. All the examples we give, as well as later work in this Chapter, are based on mood-tagged blog posts from LiveJournal; additional information about this collection and the means of obtaining it is given in Appendix B.

6.2.1 Mood Cycles

One clear observation about fluctuation of moods over time is the cyclic, regular nature of some moods. Figure 6.2 shows three examples, plotting the prevalence of three moods over time: *sleepy*, which has a clear 24-hour cycle;[3] *drunk*, with a weekly cycle (drunkenness peaks on weekends); and the yearly cycle of *stress*.[4] In the latter example, stress is low during the summer, and increases substantially throughout the autumn (as many of the bloggers are high-school or college aged [178], this correlates well with the start of the school and academic year). In the period prior to Christmas, stress increases substantially; reasons

[3] Although the blog posts in the collection come from various time zones around the globe, in 2005—the period from which most data in this Chapter is used—more than 80% of the LiveJournal users for which country information was available were from North America [178]. This means that the vast majority of posts use the same 4-5 time zones, and that skew relating to differences in zones is limited.

[4] In these and subsequent graphs of moods in the blogspace, the X-axes mark the time, and the Y-axes show the percentage of blog posts manually annotated by their authors with a given mood.

may include exams and end-of-year preparations. During the holiday period itself, stress decreases to sub-summer levels, but quickly returns to earlier levels. Note the pulsating nature of stress throughout the entire period: this is due to a secondary cycle in stress, a weekly one (stress levels decrease substantially during the weekends).

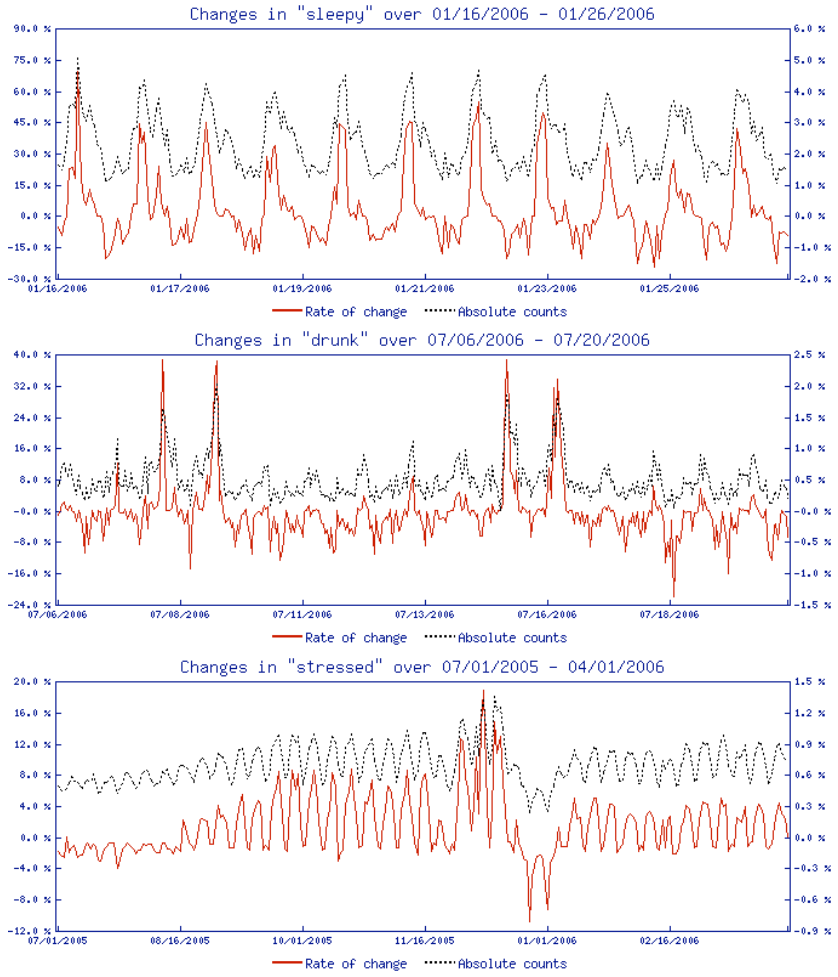

Figure 6.2: Regularities of mood fluctuation. Top: daily cycle of *sleepy* over a period of 10 days. Center: weekly cycle of *drunk* over a period of 14 days. Bottom: yearly pattern of *stress* over 8 months, slowly increasing toward the end of the year and rapidly decreasing as the new year starts. Note that the time span between tick marks is different in the different graphs.

Other moods show a less regular behavior over time. For example, Figure 6.3 shows the changes in the mood *cold*, peaking over the winter during particularly cold periods.

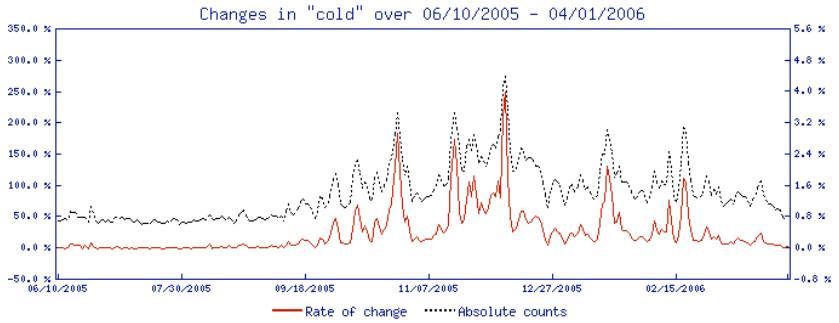

Figure 6.3: Long-term irregularities of mood behavior: *cold* peaks occasionally appear during the winter.

6.2.2 Events and Moods

Global events are clearly reflected in the mood of the blogspace. Figure 6.4 shows examples for the reaction of bloggers to two considerably major events: a peak of *sympathy* towards victims of bombings in London on July 7th, 2005, and a similar, prolonged sense of *worry* after Hurricane Katrina hit New Orleans two months later.

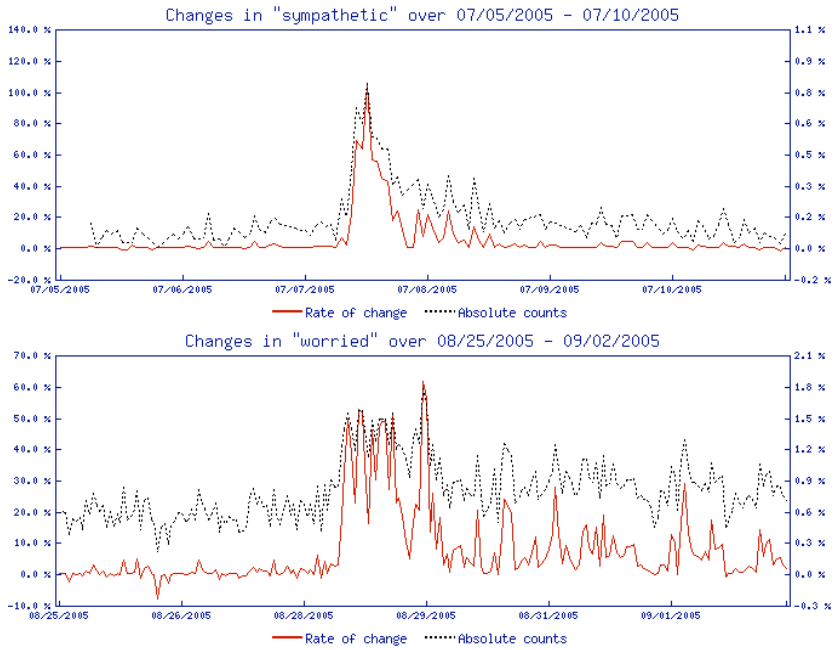

Figure 6.4: Global moods respond to global events. Top: *sympathy* spikes after bombs explode in London on July 7th, 2005. Bottom: elevated levels of *worry* as Hurricane Katrina strikes the Gulf Coast of the United States.

Placing the behavior of different moods side by side emphasizes relations between them: in a rather esoteric example, Figure 6.5 shows how a feeling of *hunger* growing as the American Thanksgiving meal approaches is replaced by a satisfied post-dinner *full* feeling.

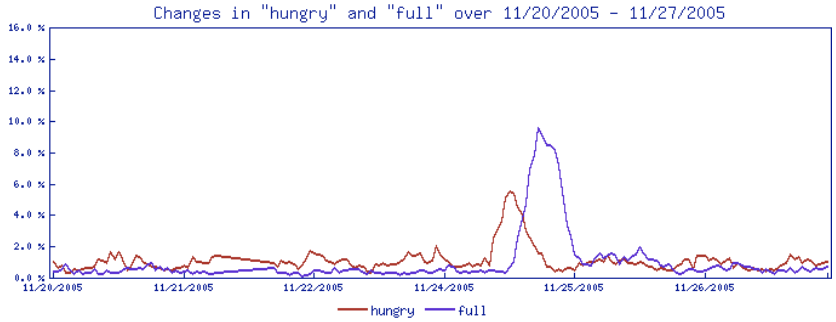

Figure 6.5: Comparing mood changes: *hungry* vs. *full* around Thanksgiving 2005.

While these may look like straightforward examples—it is clear that people are moved by tragedies and excited about holidays—they hint at the power of aggregate mood analysis. As we will see later, some mood changes are not trivial to understand without some cultural context; and even "anticipated" mood behavior can provide important observations. For example, a comparison of the level of shock or surprise of bloggers towards different events reveals their relative importance to people. Although our data comes primarily from U.S.-based bloggers, the level of sadness following the bomb attacks in London mentioned earlier was higher than the sadness level after Hurricane Katrina hit the United States; both were yet lower than the level of shock following the premature death of a television persona, Steve Irwin, in an underwater accident in September 2006.

We follow by introducing the two text analysis tasks we explore in the domain of aggregate moods in the blogspace: predicting the aggregate mood behavior from the contents of blogs, and identifying the reasons for unusual mood changes.

6.3 Predicting Mood Changes

Many blog posts are "annotated" with moods—but many, many more are not, either because the platform used does not support this annotation or because the blogger chooses not to use it. To understand the true mood of the blogspace—the one reflecting all bloggers, not only those with an explicit mood indications, we need to turn to that component of blog posts which exists for all posts—the contents of the posts. This is the task addressed in this section: not to classify the mood of individual posts, but to determine the aggregate mood levels across the entire blogspace at a given time: the intensity of "happiness" or "excitement" as

reflected in the moods of bloggers during a particular time period. The relative intensities of moods as observed in those blog posts which do include a mood indication serve as the ground truth: in a nutshell, we are trying to derive the graphs shown in the previous section from the content of blog posts (not only the manually annotated ones, but a larger set). Our research question is whether this can be done effectively and efficiently.

6.3.1 Related Work

Clearly, the work discussed here continues one of the tasks addressed in Chapter 4, namely, classifying moods of single posts. But the task we face now is different not only in the amount of data analyzed, but also in its transient nature. Moods are a fast-changing attribute, and we focus on estimating the mood levels in a certain time slot.

While research on sentiment analysis is plentiful, work on change of sentiment over time as reflected in a text corpus is scarce. A notable exception is the work of Tong on sentiment timelines, where positive and negative references to movies are tracked in online discussions over time [292]. Outside the sentiment analysis area, work on timelines in corpora is mature, mostly driven by the Topic Detection and Tracking efforts (TDT, [8]) mentioned in Section 2.3.1. Non-sentiment-related trends over time in the blogspace are exposed by some blog search engines such as BlogPulse and Technorati which provide daily counts of terms in their indices.

At the time it was made public (mid 2005), the analysis presented here was the first available work on affect behavior in blogs over time.

6.3.2 Capturing Global Moods from Text

Formally, given a set of blog posts and a temporal interval, our task is to determine the prevalence of each one of a list of given moods in the posts; evaluation is done by comparing the relative mood levels obtained from the text with the relative mood levels as reflected in the mood indicators existing in some of the posts. We approach this as a multivariate regression problem, where the response variables are the mood levels and the predictors need to be derived from the text of the posts. The task is to find the relation between these predictors and the mood intensities; we follow with details on extracting the predictors and building the regression models.

Mood predictors. Our first goal is to discover textual features that are likely to be useful in estimating prevalence of moods in a given time slot. In Section 4.1 we introduced a wide range of text-based features for classifying moods in blogs, including word and POS n-gram frequencies, special characters, PMI values, and so on. As we are now dealing with a much higher volume of text—thousands of blog posts are written every minute—we limit ourselves to the basic features,

- The hour of the day from which the data in this instance came (between 0 and 23).
- A binary indication of whether the day of the week to which this instance relates is a weekend day (i.e., Saturday or Sunday).
- The total amount of blog entries posted in this hour.
- For each discriminating term, its frequency: the percentage of blog posts containing it.

Figure 6.6: Predictors of mood intensities used for the regression analysis.

which are computationally inexpensive and easy to calculate even for large volumes of text: frequencies of word n-grams.

Rather than using the frequencies of all n-grams as predictors, we use a limited list of n-grams: those that are most likely to be associated with moods. To select which n-grams appear on this limited list, we employ the same "indicative term" extraction methods we have already used on several occasions in Chapters 3 and 4. For our current task, we are interested in terms which are indicative of multiple moods. For this, we first create lists of indicative terms for each mood we intend to predict, then merge the top terms from each of these lists.

Since our prediction is for mood intensities at a given time point, we add the time of day, as well as an indication of whether the time slot occurs on a weekend to the set of predictors. The final set of predictors is summarized in Figure 6.6.

Modeling mood levels. Once the set of predictors for mood detection is identified, we need to learn models that predict the intensity of moods in a given time slot from these predictors. Training instances are constructed from the mood-annotated data for each mood m; every training instance includes the attributes listed in Figure 6.6, as well as the "intensity" of mood m, the actual count of blog posts reported with that mood at the given time point hour.

We experimented with a number of learning methods, and decided to base our models on Pace regression [306], which combines good effectiveness with high efficiency. Pace regression is a form of linear regression analysis that has been shown to outperform other types of linear model-fitting methods, particularly when the number of features is large and some of them are mutually dependent, as is the case in our data. As with other forms of linear regression, the model we obtain for the level of mood m is a linear combination of the features, in the following format:

$$\begin{aligned} \text{MoodIntensity}_m \;=\; & \alpha_1 \cdot \text{total-number-of-posts} \;+ \\ & \alpha_2 \cdot \text{hour-of-day} \;+ \\ & \alpha_3 \cdot \text{freq}(t_1) \;+ \\ & \alpha_4 \cdot \text{freq}(t_2) \;+ \\ & \ldots, \end{aligned}$$

where t_i are the discriminating terms, and the values of α_i are assigned by the regression process.

It is important to note that both stages of our method—identifying the discriminating terms and creating models for each mood—are performed offline, and only once. The resulting models are simple, computationally cheap, linear combinations of the features; these are very fast to apply on the fly, and enable fast online estimation of "current" mood levels in the blogspace.

6.3.3 Evaluation

We now describe the experiments we performed to answer our research question—whether global mood levels can be estimated from blog text with the proposed estimation method. First, we provide details about the corpus we use; we follow with details about the discriminating terms chosen and the regression process.

Corpus. Our data consists of all public LiveJournal blog posts published during a period of 39 days, from mid-June to early-July 2005. For each entry, we store the entire text of the post, along with the date and the time of the entry. If a mood was reported for a certain blog post, we also store this indication. As described in Section 4.1, the moods used by LiveJournal users are either selected from a predefined list of 132 moods, or entered in free-text.

The total number of blog posts in our collection is 8.1 million, containing over 2.2GB of text; of these, 3.5 million posts (43%) have an indication of the writer's mood.[5]

One issue to note regarding our corpus is that the timestamps appearing in it are server timestamps—the time in which the U.S.-located server received the blog post, rather than the local time of the blogger writing the entry. While this would appear to introduce a lot of noise into our corpus, the actual effect is mild since, as we mentioned earlier, the vast majority of LiveJournal users are located in North America, sharing or nearly-sharing the time-zone of the server.

Discriminating terms. We used the text of 7 days' worth of posts to create a list of discriminating terms as described in Section 3.2; this time, we are searching for terms indicative of a specific mood rather than a given blogger. For this, we

[5] As an aside, the percentage of LiveJournal posts annotated with moods has slowly, but constantly, been decreasing since the experiments reported here; in late 2006, it was about 35%.

need to compare text associated with a given mood with more general text; in our case, the general text is the text of all posts, regardless of mood.

More specifically, for each mood m of the most popular 40 moods we aggregate the text of all posts annotated with this mood, T_m, and compare the frequencies of all unigrams and bigrams in it with their frequencies in the text of the entire corpus, T, using the log-likelihood measure described in Section 3.2. Ranking the terms by these LL values, we obtain, for each mood m, a separate list of indicative terms of that mood. This is identical to the process described when creating the indicative features for mood classification at the single blog level described in Section 4.1; examples of top-ranked indicative terms for some moods are shown in Table 6.5.[6]

Mood	**Indicative unigrams**	**Indicative bigrams**
hungry	hungry eat bread sauce	am hungry hungry and some food to eat
frustrated	n't frustrated frustrating do	am done can not problem is to fix
loved	love me valentine her	I love love you love is valentines day

Table 6.5: Most discriminating word n-grams for some moods.

Next, we combine the separate lists to a single list of indicative terms. As the amount of data we deal with is large, we aim to produce a relatively short list; the large amount of data requires a relatively simple set of features for fast analysis. For this reason, we focus only on the top-10 terms from the indicative list of each mood m; after manually filtering it to remove errors originating from technical issues (mostly tokenization problems—in total less than 10 terms were removed from all lists combined), we merge all top terms to a single list of terms indicative of moods in blogs. In total, this list contained 199 terms, of which 167 are single words and the rest word bigrams. Some examples of the discriminating terms in this list are shown in Table 6.6, along with the moods including them on their top-10 indicative term list.

Instances. The posts included in the 7 days that were used to identify the discriminating terms were removed from the corpus and not used for subsequent parts of the experiments. This left us with 32 days' worth of data for generating

[6]This is a repetition of the information in Table 4.2, and appears here for convenience.

Term	Source moods
love	cheerful, loved
envy	busy, sad
giggle	contemplative, good, happy
went to	contemplative, thoughtful
work	busy, exhausted, frustrated, sleepy, tired

Table 6.6: Examples of discriminating terms in our feature set.

the models and testing them. Instances were created by collecting, for every hour of those 32 days, all posts time-stamped with that hour, yielding a total of 768 instances. The average length of a single post in this collection is 140 words, or 900 bytes; the distribution of posts during a 24-hour period is given in Figure 6.7; each single-hour instance is therefore based on 2,500–5,500 individual posts, and represents 350K–800K words.

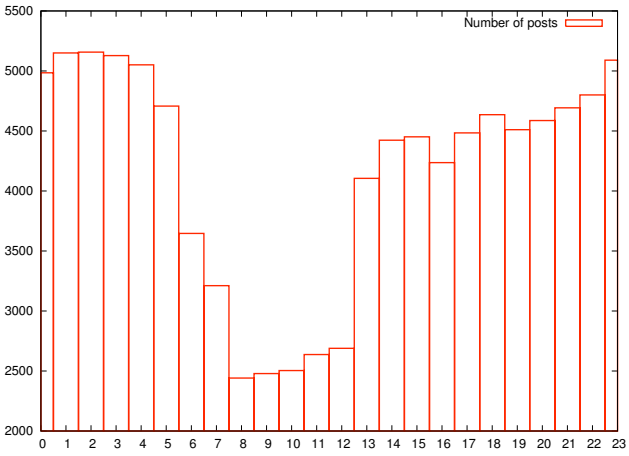

Figure 6.7: Average number of posts throughout the day. X-axis shows the hour of the day (GMT).

Generated models. We used the Pace regression module from the WEKA toolkit [314] to create our models. Since the models we create are linear regression models, they strongly exhibit the importance of features as positive and negative indicators of moods. Table 6.7 shows examples of the regression results for a couple of moods.[7]

[7]Pace regression includes a form of feature selection, therefore not all features are actually used in the resulting models.

Mood		**Linear Model**	
depressed	=	0.0123 · total-number-of-posts	+
		-523.777 · freq("accomplished")	+
		-367.5239 · freq("confront")	+
		-88.5883 · freq("crazy")	+
		-52.6425 · freq("day")	+
		90.5834 · freq("depressed")	+
		154.3276 · freq("die")	+
		-50.9185 · freq("keep")	+
		-147.1118 · freq("lol")	+
		-1137.6272 · freq("old times")	+
		283.2972 · freq("really sick")	+
		-235.6833 · freq("smoke")	+
		59.3897 · freq("today")	+
		195.8757 · freq("tomorrow")	+
		552.1754 · freq("violence")	+
		81.6886 · freq("went")	+
		-118.8249 · freq("will be")	+
		191.9001 · freq("wish")	+
		-19.23	
sick	=	-0.046 · hour-of-day	+
		0.0083 · total-number-of-posts	+
		20.3166 · freq("cold")	+
		-287.3355 · freq("drained")	+
		-91.2445 · freq("miss")	+
		-196.2554 · freq("moon")	+
		-67.7532 · freq("people")	+
		357.523 · freq("sick")	+
		615.3626 · freq("throat")	+
		60.9896 · freq("yesterday")	+
		1.6673	

Table 6.7: Examples of mood level models.

Experiments. All 768 instances of data were used to perform a 10-fold cross-validation run. The performance measures we use for our estimation are *correlation coefficient* and *relative error*. The *correlation coefficient* is a standard measure of the degree to which two variables are linearly related, and is defined as

$$\text{CorrCoefficient} = \frac{S_{PA}}{S_P \cdot S_A},$$

where

$$S_{PA} = \frac{\Sigma_i (p_i - \overline{p}) \cdot (a_i - \overline{a})}{n-1}$$

$$S_P = \frac{\Sigma_i (p_i - \overline{p})^2}{n-1}, \quad S_A = \frac{\Sigma_i (a_i - \overline{a})^2}{n-1},$$

and p_i is the estimated value for instance i, a_i is the actual value for instance i, $\overline{x}$ is the average of x, and n is the total number of instances. The *relative error* denotes the mean difference between the actual values and the estimated ones, and is defined as:

$$\text{RelError} = \frac{\Sigma_i (|p_i - a_i|)}{\Sigma_i (|a_i - \overline{a}|)}.$$

The correlation coefficient indicates how accurate the mood estimation is *over time*, showing to what degree the fluctuation patterns of a mood are predicted by the model. This is our primary metric, since we view estimation of the mood's behavior over time (e.g., detection of peaks and drops) as more important than the average accuracy as measured at each isolated point in time (which is given by the relative error). A correlation coefficient of 1 means that there is a perfect linear relation between the prediction and the actual values, whereas a correlation coefficient of 0 means that the prediction is completely unrelated to the actual values.[8]

As a baseline, we perform regression on the non-word features only, i.e., the hour of the day, the total amount of posts in that hour, and whether the day is a weekend day or not. As we demonstrated in the previous section, many moods display a circadian rhythm; because of this, and the strong dependence on the total amount of moods posted in a time slot, the baseline already gives a fairly good correlation for many moods (but the error rates are still high).

Table 6.8 shows the results of our experiments for the 40 most frequent moods. The relative high relative error rates reiterate our findings from Chapter 4 regarding the difficulty of the mood classification task itself; isolated from the temporal context, the accuracy at each point of time is not high. However, the high correlation coefficients show that the temporal behavior—the fluctuation over time—is

[8] More generally, the square of the correlation coefficient is the fraction of the variance of the actual values that can be explained by the variance of the prediction values; so, a correlation of 0.8 means that 64% of the mood level variance can be explained by a combination of the linear relationship between the prediction, and the acutal values and the variance of the prediction itself.

well predicted by the model. Also shown in Table 6.8 are the improvements of the regression over the baseline: in almost all cases the correlation coefficient increased and the relative error decreased, with substantial improvements in many cases. Note that the range of the changes is quite broad, both for the correlation coefficient and for the relative error. The average and median increase in correlation coefficient are 19% and 5.3%, respectively, and the average and median decrease in relative error are 18% and 9.7%, respectively.

The correlation levels achieved are fairly high: to demonstrate this, Figure 6.8 shows the estimated and actual levels of the mood *good* over a period of 6 days, with a correlation of 0.84—slightly higher than the average correlation achieved for all moods, 0.83.[9]

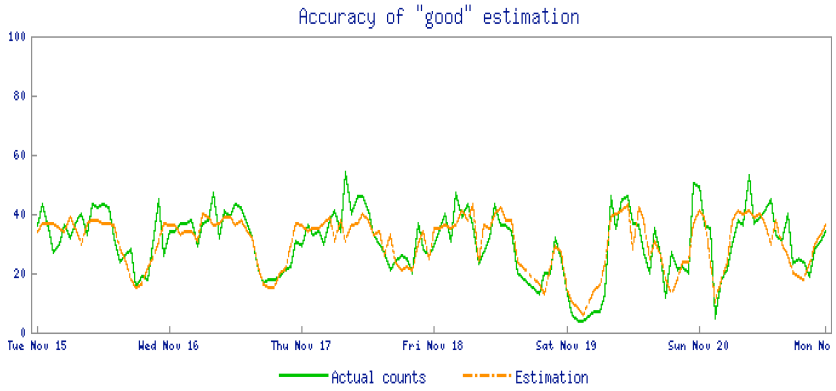

Figure 6.8: Example of mood prediction over time: prevalence of *good* (green, solid line) and the estimation based on the proposed method (orange, dashed line); the correlation in this case is 0.84.

What causes the difference in performance of our estimator across different moods? One hypothesis could be that moods for which our estimator scores higher (e.g., "bored," "happy") tend to be expressed with a small number of fairly specific words, whereas moods on which our estimator scores lower (e.g., "cold," "touched") are associated with a far broader vocabulary.

6.3.4 Case Studies

We now present two particular test cases, exhibiting particular mood prediction patterns. For these test cases, we divided our 32-day corpus into two parts: just over 24 days (585 hours) during June 2005, and just over 7 days (183 hours)

[9]This graph was created using MoodViews, a tool based on the method presented in this section and described in more detail in Appendix B.

	Correlation Coefficient			Relative Error		
Mood	**Baseline**	**Regression**	**Change**	**Baseline**	**Regression**	**Change**
drunk	0.4070	0.8611	+111.57%	88.39%	53.20%	−39.81%
tired	0.4882	0.9209	+88.63%	88.41%	37.09%	−58.04%
sleepy	0.5157	0.9106	+76.57%	80.46%	39.46%	−50.94%
busy	0.5346	0.8769	+64.02%	82.46%	45.15%	−45.24%
hungry	0.5601	0.8722	+55.72%	78.56%	44.06%	−43.91%
angry	0.5302	0.7944	+49.83%	73.70%	70.13%	−4.84%
exhausted	0.6212	0.9132	+47.00%	77.68%	39.32%	−49.38%
scared	0.4457	0.6517	+46.21%	80.30%	84.07%	+4.70%
distressed	0.5070	0.6943	+36.94%	77.49%	76.95%	−0.69%
sad	0.7243	0.8738	+20.64%	55.53%	49.91%	−10.12%
excited	0.7741	0.9264	+19.67%	61.78%	36.68%	−40.62%
horny	0.6460	0.7585	+17.41%	75.63%	63.44%	−16.11%
bored	0.8256	0.9554	+15.72%	54.22%	26.08%	−51.89%
drained	0.7515	0.8693	+15.67%	65.51%	49.50%	−24.44%
cold	0.5284	0.5969	+12.96%	87.02%	82.94%	−4.69%
depressed	0.8163	0.9138	+11.94%	57.45%	39.47%	−31.28%
anxious	0.7736	0.8576	+10.85%	60.02%	49.67%	−17.23%
loved	0.8126	0.8906	+9.59%	57.86%	44.88%	−22.43%
cheerful	0.8447	0.9178	+8.65%	50.93%	37.67%	−26.04%
chipper	0.8720	0.9212	+5.64%	47.05%	37.47%	−20.36%
bouncy	0.8476	0.8924	+5.28%	50.94%	41.31%	−18.9%
satisfied	0.6621	0.6968	+5.24%	72.97%	70.42%	−3.50%
sick	0.7564	0.7891	+4.32%	64.00%	60.15%	−6.01%
thankful	0.6021	0.6264	+4.03%	78.07%	77.48%	−0.75%
okay	0.8216	0.8534	+3.87%	54.52%	50.23%	−7.86%
ecstatic	0.8388	0.8707	+3.80%	52.35%	47.27%	−9.71%
amused	0.8916	0.9222	+3.43%	43.55%	37.53%	−13.8%
aggravated	0.8232	0.8504	+3.30%	54.91%	50.32%	−8.36%
touched	0.4670	0.4817	+3.14%	86.11%	85.39%	−0.83%
annoyed	0.8408	0.8671	+3.12%	52.28%	48.30%	−7.61%
thoughtful	0.7037	0.7251	+3.04%	69.38%	67.83%	−2.23%
crazy	0.8708	0.8932	+2.57%	46.87%	42.84%	−8.58%
cranky	0.7689	0.7879	+2.47%	63.01%	60.89%	−3.36%
happy	0.9293	0.9519	+2.43%	34.72%	28.86%	−16.86%
calm	0.8986	0.9146	+1.78%	41.89%	38.20%	−8.81%
curious	0.7978	0.8110	+1.65%	57.30%	55.69%	−2.82%
hopeful	0.8014	0.8139	+1.55%	58.79%	57.40%	−2.37%
good	0.8584	0.8714	+1.51%	51.30%	48.86%	−4.75%
optimistic	0.5945	0.6024	+1.32%	80.60%	80.25%	−0.44%
confused	0.8913	0.9012	+1.11%	44.96%	42.99%	−4.37%
average	0.7231	0.8320	+18.92%	63.56%	51.67%	−17.69%

Table 6.8: Mood level estimation for the 40 most frequent moods: 10-fold cross-validation over data from 32 days.

during July 2005.[10] The 24-day period was used for creating models, and the 7-day period for the actual case studies.

Terror in London. On the 7th of July 2005, a large-scale terror attack took place in London, killing dozens and wounding hundreds; this attack was strongly reflected in the mass media during that day, and was also a primary topic of discussion for bloggers. Following the attack, the percentage of bloggers reporting moods such as "sadness" and "shock" climbed steeply; other moods, such as "amused" and "busy," were reported with significantly lower levels than their average. An example of the reaction of bloggers can be seen in Figure 6.4 in the previous section.

Our method failed to predict both of these phenomena: the rise of negative moods and the fall of positive ones. Figure 6.9 shows two examples of the failure, for the moods "sadness" and for "busy." The correlation factors for some moods, such as these two, drop steeply for this period.

An examination of the blog posts reported as "sad" during this day shows that the language used was fairly unique to the circumstances: recurring words were "terror," "bomb," "London," "Al-Qaeda," and so on. Since these words were not part of the training data, they were not extracted as indicative features for sadness or shock, and were not included in our estimation method.

We hypothesized that given the "right" indicative words, our method would be able to estimate also these abnormal mood patterns. To test our hypothesis, we modified our data as follows:

- Add the two words "attack," and "bomb" to the list of words used as discriminating terms. These were the top overused terms during this time period, according to the log-likelihood measure we use to identify indicative words.

- Move two instances from the test data to the training data; these two instances reflect two hours from the period of "irregular mood behavior" on July 7th (the hours selected were not the peak of the spikes).

This emulates a scenario where the language used for certain moods during the London attacks has been used before in a similar context; this is a likely scenario if the training data is more comprehensive and includes mood patterns of a larger time span, with more events.[11]

[10] These consist of July 1st to July 3rd, and July 6th to July 9th. We have no data for two days—July 4th and 5th—due to technical issues.

[11] In the particular case where there is a stream of data updated constantly, some of it annotated—as is the case with blog posts—this can be done automatically: the quality of the estimation is measured with new incoming annotated data, and when the quality drops according to some critera, the models are retrained.

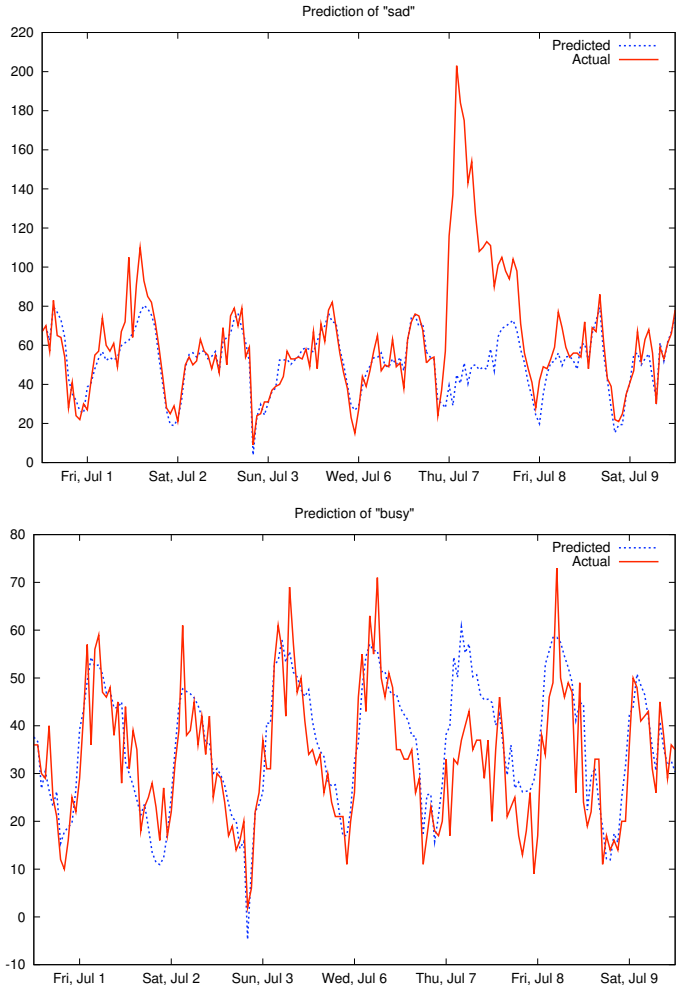

Figure 6.9: Failure to predict a sadness spike following the terror attacks in London (top), and the accompanying decrease in busyness (bottom). Counts of posts are indicated on the Y-axis; the red, continuous line marks actual counts, and the blue, dashed line is the prediction.

We then repeated the estimation process with the changed data; the results for "sadness" are shown in Figure 6.10. Accordingly, the correlation values climb back close to those achieved in our 10-fold cross-validation.

Weekend drinking habits. Our next test case is less somber, and deals with the increased rate of certain moods over weekends, compared to weekdays—already mentioned when discussing cyclic mood patterns in the previous section.

Figure 6.11 shows our estimation graphs for the moods "drunk" and "excited" for the same period as the previous London bombing test case—a period including

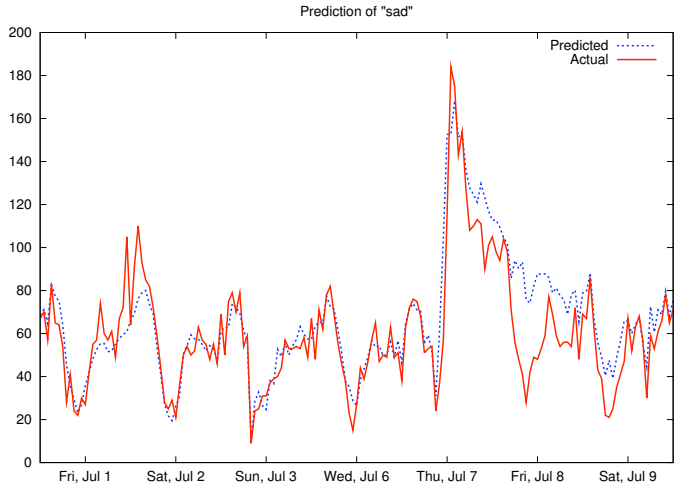

Figure 6.10: Successful prediction of the sadness peak with modified data. Counts of posts are indicated on the Y-axis.

two weekends. Clearly, both moods are successfully predicted as elevated during weekends, although not at the full intensity.

6.3.5 Conclusions

The work we presented aims at identifying the intensity of moods in the blogspace during given time intervals. Using a large body of blog posts manually annotated with their associated mood, we achieve high correlation levels between predicted and actual moods by using words which are indicative of certain moods. Our main finding is that while prediction of mood at the individual blog post level is a hard task, as shown in Section 4.1, at the aggregate level, predicting the *intensity* of moods over a time span can be done with a high degree of accurracy, even without extensive feature engineering or model tuning. Having said that, we believe that further expansions of the predictor set, i.e., using a larger amount of discriminating terms, and using any of the features used for single-post mood classification, will improve the results further.

6.4 Explaining Irregular Mood Patterns

The previous section included some examples of irregular mood behavior in the blogspace, such as peaks of certain moods after large-scale global events. In many of these cases, the reason for the irregular behavior is clear to the observer, assuming she shares the same cultural context as the bloggers. If we know that there has been a major tragedy we expect people to be shocked or sad; we assume

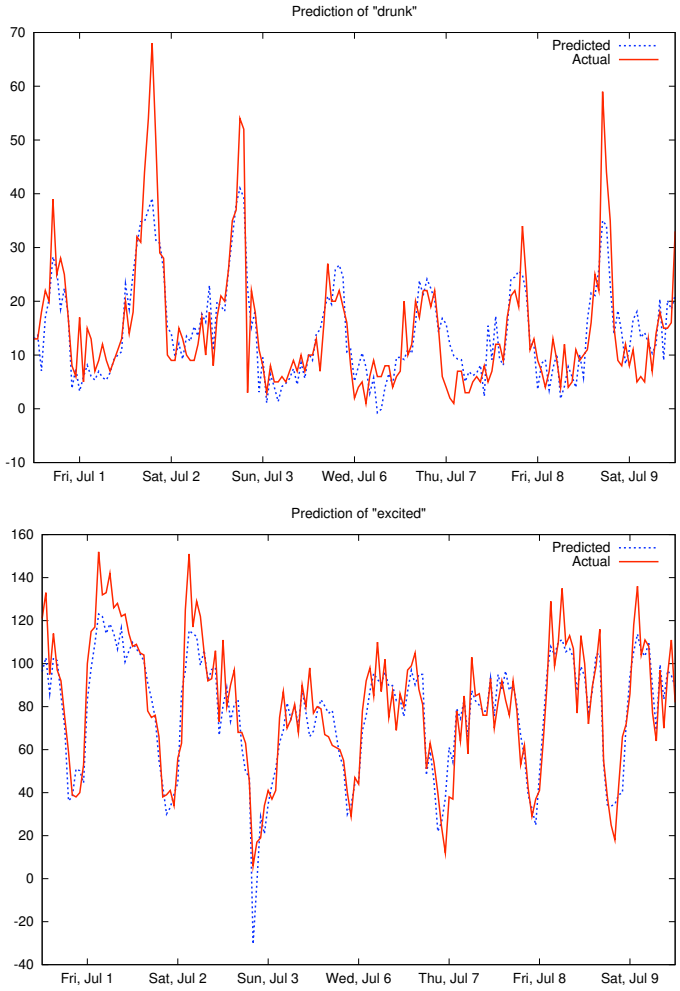

Figure 6.11: Prediction of weekend-related moods: "drunk" (top) and "excited" (bottom). Counts of posts are indicated on the Y-axis.

elevated relaxation during holidays, and so on.

However, not all irregularities in mood fluctuations are easily explained, and some require very specific context to understand. Consider, for example, the spike in excitement experienced in the blogspace in mid-July 2005 (Figure 6.12): what has happened on this day to make bloggers react so strongly? When we first encountered this peak, we were not aware of any large-scale event with such an expected effect.

In this section, we develop a method to address this and similar questions. Our approach identifies unusual changes in mood levels in blogs and locates an explanation for the underlying reasons for these changes: a natural-language text describing the event that caused the unusual mood change.

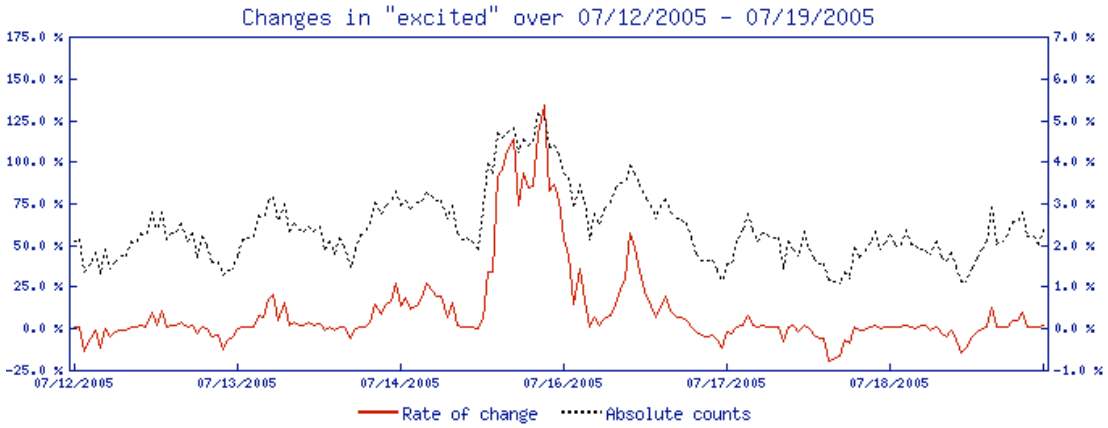

Figure 6.12: Surge in excitement in blogs on July 16th, 2005.

The method we use to produce such explanations is as follows. First, we compare the expected mood levels and the actual ones to identify irregular behavior. If unusual spikes occur in the level of mood m, we examine the language used in blog posts labeled with m around and during the period in which the spike occurs. We compare this language to the long-term language model for m, using overused terms for the irregular period as indications for the mood change. Once these terms are identified, we use them to consult a collection of global events—a news corpus—from which we retrieve a small text snippet as the desired explanation.

Our work is related to the burstiness models described by Kleinberg in time-lined corpora such as email and research papers [147]; there, irregularities are identified by applying probability models used to analyze communication networks. The same model is applied to discover dense periods of "bursty" intra-community link creation in the blogspace [156] and to identify topics in blogs over time [217].

6.4.1 Detecting Irregularities

Our first task, then, is to identify spikes in moods reported in blog posts: unusually elevated or degraded levels of a mood in a particular time period. As shown earlier, many moods display a cyclic behavior pattern, maintaining similar levels at a similar time-of-day or day-of-week (see Section 6.2). Our approach to detecting spikes addresses this by comparing the level of a mood at a given time point with the "expected" level—the level maintained by this mood during other, similar time points. Formally, let POSTS($mood, date, hour$) be the number of posts labelled with a given mood and created within a one-hour interval at a specified date. Similarly, ALLPOSTS($date, hour$) is the number of all posts created within the interval specified by the date and hour. The ratio of posts labeled with a given mood to all posts for a day of a week (Sunday, ..., Saturday) and

for a one-hour intervals (0, ..., 23) is given by:

$$R(mood, day, hour) = \frac{\sum_{\mathrm{DW}(date)=day} \mathrm{POSTS}(mood, date, hour)}{\sum_{\mathrm{DW}(date)=day} \mathrm{ALLPOSTS}(date, hour)}$$

where $day = 0, \ldots, 6$ and $\mathrm{DW}(date)$ is a day-of-the-week function that returns 0, ..., 6 depending on the date argument.

The level of a given mood is *changed* within a one-hour interval of a day, if the ratio of posts labelled with that mood to all posts, created within the interval, is significantly different from the ratio that has been observed on the same hour of the similar day of the week. Formally:

$$D(mood, date, hour) = \frac{\frac{\mathrm{POSTS}(mood,date,hour)}{\mathrm{ALLPOSTS}(date,hour)}}{R(mood, \mathrm{DW}(date), hour)}$$

If $|D|$ exceeds a threshold we conclude that an unusual spike occurred, while the sign of D makes it possible to distinguish between positive and negative spikes. The absolute value of D expresses the degree of the peak. Consecutive hours for which $|D|$ exceeds the thresholds are grouped into a single interval, where the first hour marks the start of the peak and the last one is the end of it.

6.4.2 Explaining Irregularities

Once an irregular interval is identified, we proceed to the next stage: providing an explanation to the irregular behavior. First, we identify terms which are indicative of the irregularity. For this, we follow the same language-model-based keyword extraction approach used to identify terms associated with a particular mood in the previous Section; however, we now attempt to identify indicative terms for a given mood *during a given period*, rather than terms indicative of it regardless of time. To this end, we compare the language model associated with the mood m during the irregular period with the language model associated with m in other periods, generating a ranked list of indicative terms.

After identifying a set of terms related to the event behind the unusual pattern, the next step is straightforward: the top indicative terms are matched with a set of descriptions of events that took place during a time corresponding to the irregularity. Such timestamped descriptions are easy to obtain, e.g., through newswire corpora or streamed headlines.

Due to the length and high quality of the "queries" used—each query term is typically highly-related to the event—the effectiveness of this retrieval process is high, particularly for early precision. Different, unrelated events taking place within the same short time period share little vocabulary: we found that a few query terms, and a simple ranking mechanism (measuring the overlap between the top overused terms and the title of the event) provide good results.

6.4.3 Case Studies

Evaluation of the method described here is non-trivial. Instead, we show a few test cases demonstrating its usefulness. In these examples, the corpus used is identical to the one from the previous section—public blog posts of LiveJournal, starting from July 2005. As a news corpus, we used the English edition of Wikinews (http://en.wikinews.org), a collaborative site offering syndicated, royalty-free news articles.

For the first example, we return to the unusual peak of excitement appearing in Figure 6.12—the one which was unclear to us and prompted the development of the method this Section describes. Figure 6.13 shows this peak again; this time, the interval identified as unusual by our method is highlighted. The top overused terms during this period were "harry," "potter," "book," "hbp," "excited," and "prince." The headline of the top Wikinews article retrieved for the date of the peak using these terms is "July 16: Harry Potter and the Half-Blood Prince released" (recall that the average blogger is in her teens—an age group where this particular book series is extremely popular). Clearly, this is a simple, short, and effective explanation for excitement among bloggers .

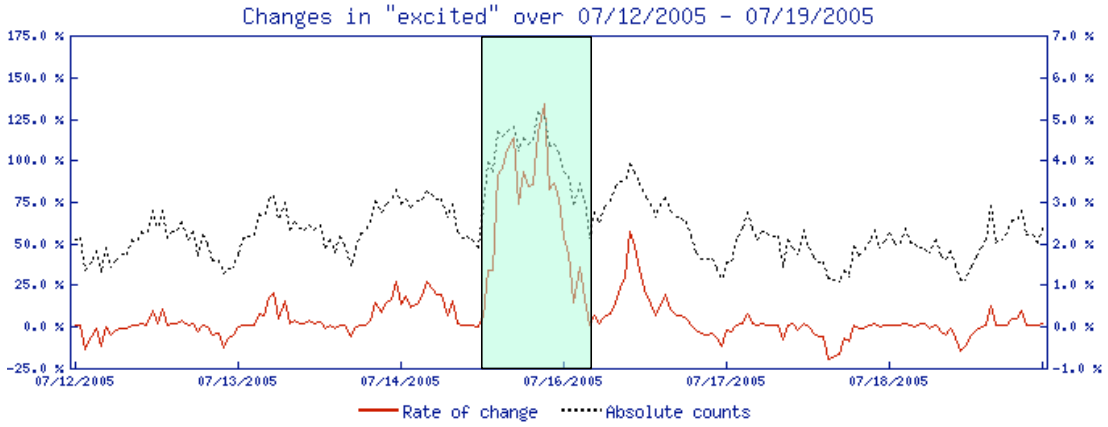

Figure 6.13: Surge in excitement in blogs on July 16th, 2005: surge interval identified, overused terms extracted, and top headline retrieved.

The next example concerns another event with a much stronger influence on bloggers than anticipated. During September 4th, 2006 (and, to a lesser extent, September 5th), significant peaks of sadness, shock and sympathy were registered, again with no obvious explanation to us as observers. The top overused terms for these moods during this time were "crocodile," "australia," "sting," "underwater," and "irwin"; the top retrieved Wikinews article for these terms described the death of Steve Irwin, the star of a television show called "The Crocodile

Hunter," in an underwater accident off the coast of Australia. Figure 6.14 shows the results of our method for this case: the highlighted peak, the overused terms extracted and the top headline; again, the resulting explanation clearly provides context for the irregularity.

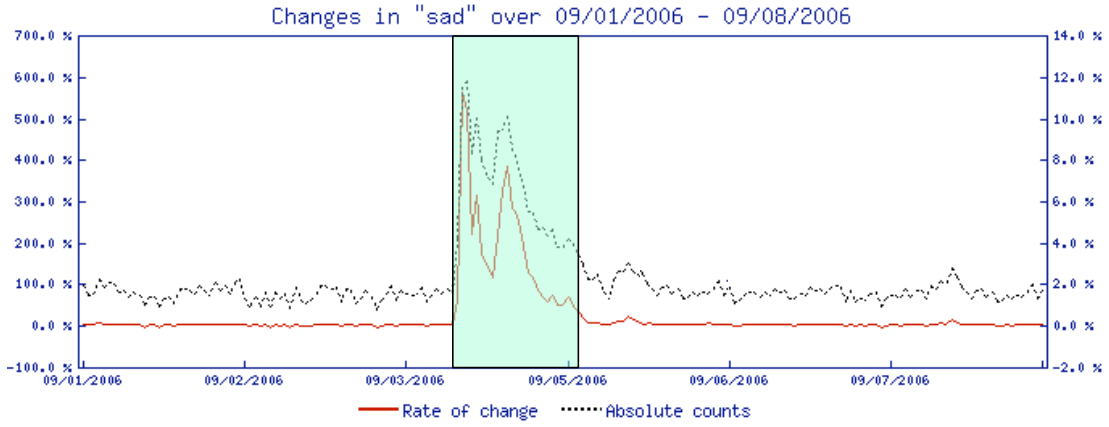

Figure 6.14: Surge in sadness in blogs on September 4th, 2006: surge interval identified, overused terms extracted, and top headline retrieved.

In addition to these two examples, the method we proposed was used to identify causes for other unusual mood phenomena including those shown on Figure 6.4, caused by major news events. More details about a web service built around this method are provided in Appendix B.

6.4.4 Conclusions

We described a method for relating changes in the global mood, as reflected in the moods of bloggers, to large-scale events. The method is based on identifying changes in the vocabulary bloggers use over time, and, particularly, identifying overused terms during periods in which the reported moods by bloggers differ substantially from expected patterns. Overused terms are then used as queries to a collection of time-stamped events, resulting in "annotations" of irregularities in global moods with human-readable explanations. While rigorous evaluation of such a task is complex, anecdotal evidence suggests that the resulting explanations are useful, and answer the question we set out to explore earlier in this Section.

More generally, the process of selecting overused terms during a specific time period in a timelined corpus (and, possibly, matching them with a corpus of events) can be used in other, non-mood-related scenarios. For example, it can

be used to annotate peaks in occurrences of terms in a corpus over time,[12] or to track the changes in interests and opinions of a blogger towards a given topic by following the changing language models associated with this topic in her blog.

To summarize, this Chapter focused on aspects of aggregate sentiment analysis in blogs. First, we motivated this work by demonstrating its usefulness: we show that for analysis of Consumer Generated Media, sentiment analysis provides better prediction of financial success of products than the volume of discussion only, and discussed the relation between the accuracy of the prediction and the level of context used for the sentiment classification. We then demonstrated that the collective mood reports of bloggers provide insight into more than commercial data only, namely, into global behavior patterns. This latter observation has led to two research questions, explored in the subsequent parts of the Chapter. The first was whether the global mood can be inferred from the aggregate text of bloggers. To answer this, we coupled a regression-based approach with indicative term mining, showing that the answer is positive: global moods can be approximated with high degrees of accuracy using simple methods. This demonstrated the power of using large amounts of data for this task: as shown in Chapter 4, a similar task at the level of single posts resulted in substantially less accurate predictions. The second question we asked was whether irregularities in global mood behavior can be explained in an automated manner by monitoring the language use in the blogspace. Here, too, the answer is positive; to provide such explanations, we offer a method linking temporal changes in discriminative terms used by bloggers reporting a certain mood to a corpus of global events.

Taking a step back, we started with a somewhat traditional task for sentiment analysis: finding the relation between consumer opinion and product success, showing that known sentiment analysis methods are more beneficial in this domain than in others. The work discussed in the rest of the Chapter involved more novel information needs and tasks, demonstrating the type of knowledge that can be found in the blogspace, as a unique collection of people's emotions. Global mood patterns are one type of such knowledge (and a particularly singular one); we will return to non-factual aspects of blog content in Chapter III, where we address the task of locating sentiment in a large collection of blogs.

[12] Google Trends (www.google.com/trends) is an application which appears to do this for occurrences of terms in Google's search log; it was introduced after the work presented here was made public.