Capturing Global Mood Levels using Blog Posts

Gilad Mishne and **Maarten de Rijke**
ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
gilad,mdr@science.uva.nl

Abstract

The personal, diary-like nature of blogs prompts many bloggers to indicate their mood at the time of posting. Aggregating these indications over a large amount of bloggers gives a "blogosphere state-of-mind" for each point in time: the intensity of different moods among bloggers at that time. In this paper, we address the task of estimating this state-of-mind from the text written by bloggers. To this end, we build models that predict the levels of various moods according to the language used by bloggers at a given time; our models show high correlation with the moods actually measured, and substantially outperform a baseline.

Introduction

The blogosphere, the fast-growing totality of weblogs or blog-related webs, is a rich source of information for marketing professionals, social psychologists, and others interested in extracting and mining opinions, views, moods, and attitudes. Many blogs function as an online diary, reporting on the blogger's daily activities and surroundings; this leads a large number of bloggers to indicate what their mood was at the time of posting a blog entry. The collection of such mood reports from many bloggers gives a "blogosphere state-of-mind" for each point in time: the intensity of different moods among bloggers at that time. Several applications of this idea are currently being explored, in some cases commercially:

- Tracking the public affect toward certain products, brands, or people: a company may be interested in the mood reported in texts relating to its products; PR offices representing entertainment figures would be interested in the mind-sets associated with their clients.
- Discovering global mood phenomena: political scientists, as well as media analysts, have an ongoing interest in public opinions and moods, particularly the effect of policies and events on it. Automatically determining the mood associated with a piece of text enables them to access a much higher volume of data than the amount that can be analyzed manually.

This motivates the development of algorithms for estimating the general levels of moods of blog posts. The task is not to classify each individual post with respect to its mood, which has been shown to be a difficult task (Mishne, 2005). Instead, we are interested in determining aggregate mood levels across a large number of postings. That is, in a given time slot, we want to estimate the global intensity of moods such as "happy" or "excited," as reflected in the mood of bloggers. We demonstrate this using the graph in Figure 1. This graph shows the number of bloggers reporting their mood as "pleased", during a period of 5 days in August 2005, measured hourly (the details of the corpus from which this was taken are given later). We take this to reflect the "true" level of this mood within the blogosphere, and the task we set ourselves in this paper is to predict this graph without explicitely using the moods as reported by the bloggers. More formally, given a set of blog posts in some temporal interval, and given a mood, our task is to determine the fraction of the posts with that mood.

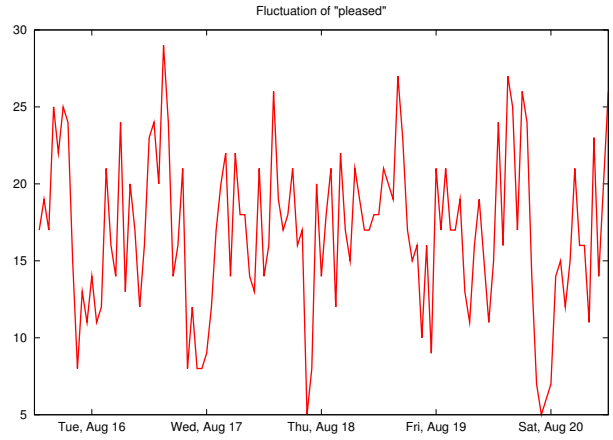

Figure 1: Number of blog posts tagged with the mood "pleased" by the authors during a 5-day interval in August 2005.

This may be viewed as a text classification task, but it differs from other text classification tasks such as determining the topic or genre of the text, or the gender of its author. The most important aspect that sets our task apart is its transient nature. Blog posts are highly related to the date and time of

their publication: often they comment on current events, on the blogger's personal life at a given moment of the day, and so on. Moreover, moods (unlike topicality or author gender) are a fast-changing attribute; happiness can quickly turn to a more relaxed state, and tiredness is in most cases a temporary state. As a result, we focus on estimating the mood levels *in a certain time slot*, rather than estimating moods of complete blogs.

The main contributions of this paper are as follows.

- A description of a mood estimation task, where indications of moods from a large amount of "real people" serve as the ground truth.
- A method for online, fast estimation of mood levels using the text of blog posts; this method substantially improves over a baseline.

The remainder of the paper is organized as follows. In the next section we discuss related work. Then, we describe our method for estimating mood levels. We evaluate our method, and before concluding we zoom in on two particular test cases.

Related Work

Recent years have witnessed an increase in research on recognizing and gathering subjective and other non-factual aspects of textual content, much of it driven by interest in consumer or voters' opinions. Sentiment analysis, i.e., classifying opinion texts or sentences as positive or negative, goes back a long time. Work of Hearst (1992) on sentiment classification of entire documents uses cognitive models. Lexicon-based methods for subjectivity classification have received a lot of attention (Das & Chen, 2001, Hatzivassiloglou & Wiebe, 2000, Kamps, Marx, Mokken, & de Rijke, 2004), especially early on, while the interest in data-driven methods has been growing rapidly in recent years, as is witnessed by research into both supervised methods (Dave, Lawrence, & Pennock, 2003, Pang, Lee, & Vaithyanathan, 2002) and unsupervised methods (Turney, 2002). Recently, a formal metric for polarity levels has been proposed (Nigam & Hurst, 2004), based on a probabilistic model.

In this paper we deal with mood developments at the aggregate level. Much of the work cited so far deals with subjectivity at the level of individual documents, even though there is some work on tracking subjectivity developments at the aggregate level on resources other than blogs. For instance, Tong (2001) generates sentiment (positive and negative) timelines by tracking online discussions about movies over time. And Liu, Hu, & Cheng (2005) aggregate features commented by customers on online reviews and what customers praise or complain about; they also perform opinion comparisons.

While there is a fair number of search engines specializing in blogs nowadays, there is relatively little research on extracting or tracking opinions or moods on blogs. 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 (Mishne, 2005), confirming, on blogs and moods, the findings of Dave et al. (2003) on product reviews and subjectivity. Related to our work, but different, are the activity and trend watching services that search engines such as BlogPulse provide (Glance, Hurst, & Tomokiyo, 2004).

Mood Tracking

We now describe the method we use for estimating mood levels in the blogosphere based on the language used by bloggers. Our estimation process is composed of two stages:

- Identifying textual features that can be used to estimate mood prevalence.
- Learning models that predict the intensity of moods in a given time slot, utilizing these features.

We proceed by providing details about these stages separately.

Discriminating Terms

As noted, our first goal is to discover features that are likely to be useful in creating models that predict mood levels. While a wide range of such features exists, including word and word n-gram frequencies, special characters, post length etc. (Mishne, 2005), we focus on the most widely-used set of features in text classification systems, namely frequencies of word n-grams in the text (Sebastiani, 2002). Our goal in this stage, then, is to identify words and phrases that are likely to indicate certain moods.

To do this, we need an annotated corpus: a collection of texts, each tagged with its author's mood. Typically, this is a difficult resource to obtain. However, in the particular case of blogs, we can rely on the unique feature of blogs mentioned earlier on, namely, the fact that bloggers often supply the mood they are in at the time of posting a blog entry. This provides us with the required body of manually-classified text; the popularity of blogging ensures a high volume of data. While the "annotators"—the bloggers—are not consistent and certainly do not follow guidelines for tagging their posts, our working assumption is that the amount of data makes up for the high level of noise in it.

Viewing our corpus as a collection of text tagged with moods enables us to identify the words and phrases that are indicative of these moods by applying existing methods for quantifying the divergence between term frequencies across different corpora. We make use of one such measure – log likelihood (Rayson & Garside, 2000). The corpora we are comparing are the text known to be associated with a certain mood, and all texts known to be associated with other moods.

More formally, for each mood m we define two probability distributions, Θ_m and $\Theta_{\overline{m}}$, to be the distribution of all words in the combined text of blog posts reported with mood m, and the distribution of all words in the rest of the blog text, respectively. We then rank all the words in Θ_m, according to their log likelihood measure, as compared with $\Theta_{\overline{m}}$: this gives us a ranked list of "characteristic terms" for mood m. Once this process has been carried out for all moods, we create a single feature set of "discriminating terms" by selecting the terms that appear in the top-N of the largest

- 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.
- The actual count of posts reported with mood m at this hour (this is the number to be estimated, the "mood intensity").

Figure 2: Attributes of instances.

number of the separate ranked lists (this was done to limit the total number of features; ideally, all top-N words should be used).

The above process can be repeated for identifying characteristic bigrams, trigrams, or higher-order n-grams.

Modeling Mood Levels

Once a set of indicative features for mood detection is identified, we need to formulate an effective way of estimating moods of bloggers based on these features. As mentioned earlier, moods are transient, and what we are actually predicting is the *level* of certain moods in a given time frame—the amount of "happiness" or "sadness" at a specific hour. For this, we group the blog posts according to their timestamps—all posts from the same hour are aggregated. Then, for each mood, we count the number of blog posts associated with each mood, as well as the number of blog posts containing each one of the discriminating terms. Finally, from this data we construct training instances for each mood m; every training instance includes the attributes listed in Figure 2. The training instances are then fed to a learning algorithm, creating, for each mood, a model of the relation between the values of the features and the intensity of the mood.

We experimented with a number of learning methods, and decided to base our models on Pace regression (Wang & Witten, 1999), which combines good effectiveness with high efficiency. Pace regression is a form of linear regression analysis that has 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:

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

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 blogosphere.

Evaluation

In this section we describe the experiments we performed to test our estimation method. First, we provide details about the corpus we use to test our prediction method; we follow with details about the discriminating terms chosen and the estimation experiments themselves.

Corpus

Our data consists of all public blog posts published in LiveJournal—the largest online blogging community[1]—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. LiveJournal users have an option to indicate a (single) mood when adding an entry; if a mood was reported for a certain blog post, we also store this indication. The moods used by LiveJournal users are either selected from a predefined list of 132 moods, or entered in free-text; for more information, see (Mishne, 2005).

One important restriction of our corpus is that it does not constitute a representative sample of the adult population; it does not even reflect a representative sample of all bloggers: most LiveJournal users are under the age of 20, there are more females than males, and so on; see (LiveJournal Statistics). Another 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 the vast majority of LiveJournal bloggers are located in the U.S. and Canada, sharing or nearly-sharing the time-zone of the server.[2] Finally, the available "moods" for LiveJournal bloggers do not correspond to any well-known model of moods or emotions such as Plutchik's wheel model or Shaver's taxonomy (Santrock, 2000); while most of these moods would generally be accepted as genuine moods (e.g., "depressed", "excited"), others are arguably not real moods ("hungry", "cold"). These are all significant shortcomings of the corpus, but given the difficulty of obtaining realistic large-scale texts with mood indications, our corpus still constitutes an excellent and unique data source.

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

[1] URL: http://www.livejournal.com

[2] In September 2005, more than 80% of the LiveJournal users for which country information is available are from the U.S. or Canada.

Discriminating Terms

We used the text of 7 days' worth of posts to create a list of discriminating terms as described in the previous section; our terms consist of the most popular single- and double-word expressions in the top-10 of the log-likelihood-ranked list of each mood. This list was manually filtered to remove some errors originating from technical issues, mostly tokenization problems; in total less than 10 terms were removed). The final list of features contains 199 terms, of which 167 are single words and the rest two-word phrases. Some examples of the discriminating terms we ended up with are given in Table 1.

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

Table 1: Examples of discriminating terms in our feature set.

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 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 number of words in a single post is 140 (900 bytes); the distribution of posts during a 24-hour period is given in Figure 3; each single-hour instance is therefore based on 2500–5500 individual posts, and represents 350K–800K words.

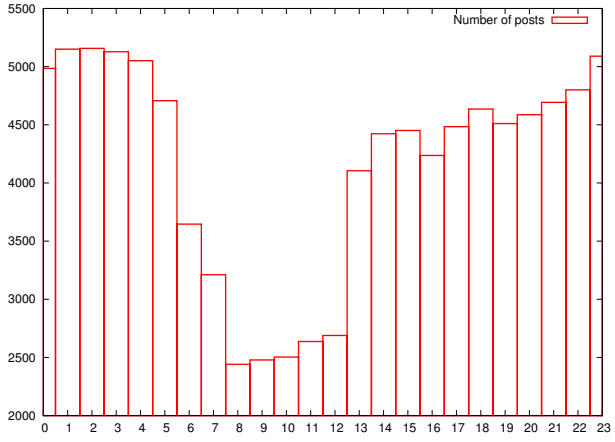

Figure 3: Average number of posts throughout the day. X-axis shows the hour of the day (GMT).

Generated Models

We used the Pace regression module from the WEKA toolkit (Witten & Frank, 2005) 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 2 shows examples of the regression results for a couple of moods.[3]

Experiments

We used all 768 instances of data 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

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

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

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.[4]

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 days is a weekend day or not. 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).

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

[4] 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.

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 2: Examples of mood level models.

Table 3 shows the results of our experiments for the 40 most frequent moods, with an indication of the improvements of the results over the baseline: in almost all cases the correlation coefficient increased and the relative error decreased; improvements were substantial in many cases. Observe 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 is 19.60% and 5.46%, respectively, and the average and median decrease in relative error is 17.12% and 9.26%, respectively.

What causes the difference in performance of our predictor 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; anecdotal evidence does not support this, though.

Case Studies

We now present two particular test cases, exhibiting our 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) during July 2005.[5] 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.

Our method failed to predict both of these phenomena: the rise of negative moods and the fall of positive ones. Figure 4 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: repeating 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:

- Manually add the two words "attack", and "bomb" to the list of words used as discriminating terms.
- 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.[6]

We then repeated the estimation process with the changed data; the results for "sadness" are shown in Figure 5. Accordingly, the correlation values climb back close to those achieved in our 10-fold cross-validation.

A useful direction to explore, for enriching the vocabulary of mood-indicative words on the fly, is analyzing a stream of news-articles which are published at the time of the blog posts, and using key words and phrases appearing there.

[5] 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.

[6] 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.

Mood	Correlation Coefficient			Relative Error		
	Baseline	**Regression**	**Change**	**Baseline**	**Regression**	**Change**
drunk	0.407	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	0.507	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	(+3.8%)	52.35%	47.27%	(−9.71%)
amused	0.8916	0.9222	(+3.43%)	43.55%	37.53%	(−13.8%)
aggravated	0.8232	0.8504	(+3.3%)	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	0.7158	0.8272	(+19.60%)	64.02%	52.53%	(−17.12%)

Table 3: Mood level estimation for the 40 most frequent moods: 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.

Figure 6 shows our estimation graphs for the moods "drunk" and "excited" for the same period as the one discussed in the previous test case—a period containing two weekends. Clearly, both moods are successfully predicted as elevated during weekends, although not at the full intensity.

As an aside, it is interesting to note the "mirror-like" appearance of these two moods: while excitement is usually reported at daytime, drunk blog entries tend to be posted at later hours.

Conclusions

The work we presented aims at identifying the intensity of moods within the blogging community during given time intervals. We use a large body of blog posts manually annotated (by the bloggers themselves) with their associated mood. Using this annotation, we identify words which are indicative of certain moods, then learn linear models for estimating the mood levels using the frequencies of the words in blog posts, as well as meta-information about the time interval itself. Our models exhibit a strong correlation with the actual moods reported by the bloggers, and significantly improve over a baseline.

Our main finding is this. While it was known that determining the mood associated with an individual blog post is a very hard task, mainly due to the limited length of posts

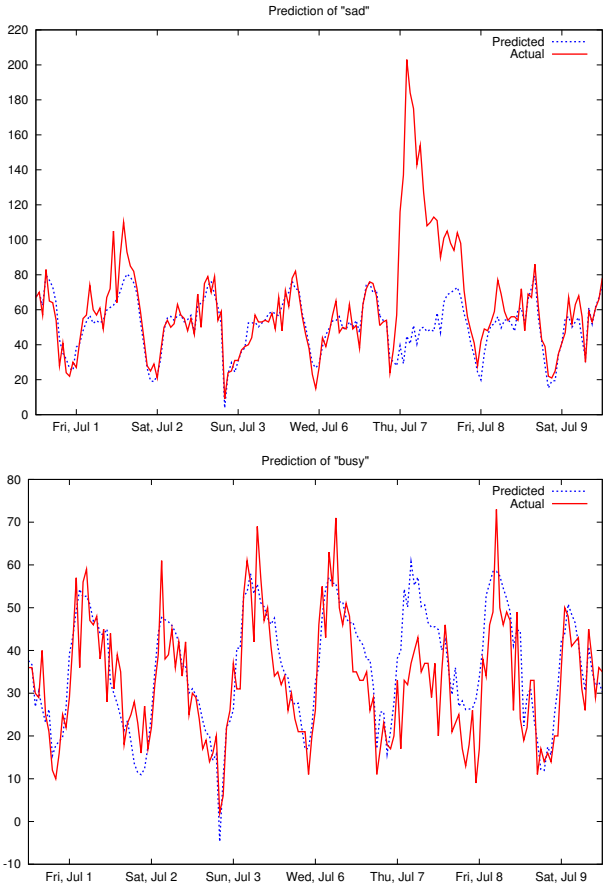

Figure 4: Failure to predict a sadness spike following the terror attacks in London (top), and the accompanying decrease in busyness (bottom). Counts of posts are indicated on the Y-axis.

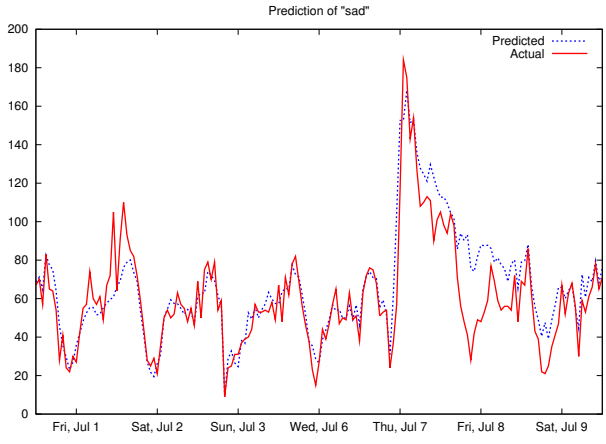

Figure 5: Successful prediction of the sadness peak with modified data. Counts of posts are indicated on the Y-axis.

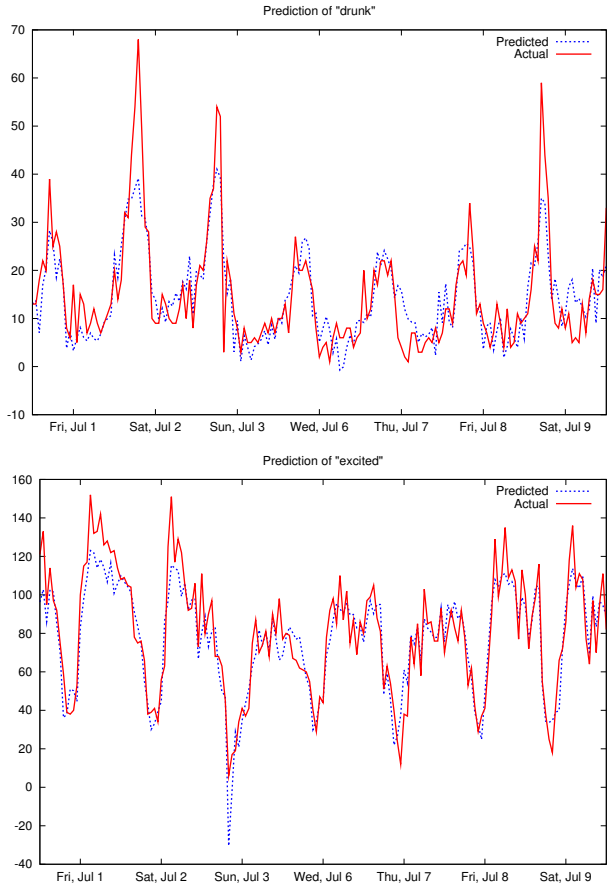

Figure 6: Prediction of weekend-related moods: "drunk" (top) and "excited" (bottom). Counts of posts are indicated on the Y-axis.

and the lack of an annotation regime, we have shown that, 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.

An online version, demonstrating our mood tracking and estimation work, is available at `http://moodviews.com/Moodteller`.

Future Directions

In addition to obvious expansions of our feature set, i.e., using a larger amount of discriminating terms, additional non-content attributes can be used for mood level prediction. Features which seem promising are the use of emoticons—textual representations of facial expressions—as well as sentiment values of the individual words, and other features.

Our feature set included an indication of the day of the week; a larger corpus, spanning months or even years, may also include indications of the month or the season, to measure their influences (e.g., winter is traditionally associated with depression, spring with joy, and so on).

Furthermore, there is obvious research that be done by revisiting some of the choices made in this paper. E.g., dif-

ferent ways of identifying discriminating terms, or using different regression methods for estimating mood levels.

Aside from the prediction task that we addressed in this paper, the unique corpus consisting of a large body of time-stamped, personally-oriented texts tagged with moods gives rise to a wealth of other interesting tasks. For example, we conducted a small-scale experiment measuring the correlation between the temporal behavior of certain moods (the "mood graph") and the temporal behavior of word frequencies in the text ("word graphs", measuring the occurrences of words over time). Anecdotal evidence shows that this correlation is meaningful, e.g., the word "happy" is the highest-correlating word with the mood "loved". In the same manner, it is possible to cluster moods or terms according to their temporal behavior.

Another direction we are exploring concerns methods for automatically associating current events with unusual behavior in specific moods.

Acknowledgments We thank Matthew Hurst and Natalie Glance for valuable comments, and our referees for helpful suggestions.

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,

References

S. Das & M. Chen. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings APFA 2001*, 2001.

K. Dave, S. Lawrence, & D. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings WWW 2003*, 2003.

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

V. Hatzivassiloglou & J. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings COLING 2000*, 2000.

M. Hearst. Direction-based text interpretation as an information access refinement. In P. Jacobs, editor, *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, pages 257–274. Lawrence Erlbaum Associates, Hillsdale, 1992.

J. Kamps, M. Marx, R. Mokken, & M. de Rijke. Using WordNet to measure semantic orientations of adjectives. In *Proceedings LREC 2004*, volume IV, pages 1115–1118, 2004.

B. Liu, M. Hu, & J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings WWW 2005*, pages 342–351, 2005.

LiveJournal Statistics. LiveJournal Statistics. URL: http://www.livejournal.com/stats.bml, 2005. Accessed July 2005.

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.

K. Nigam & M. Hurst. Towards a robust metric of opinion. In *The AAAI Symposium on Exploring Attitude and Affect in Text (AAAI-EAAT)*, 2004.

B. Pang, L. Lee, & S. Vaithyanathan. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In *Proceedings EMNLP 2002*, 2002.

P. Rayson & R. Garside. Comparing corpora using frequency profiling. In *The workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*, 2000.

J. W. Santrock. *Psychology*. McGraw-Hill, 2000.

F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002. ISSN 0360-0300.

R. Tong. An operational system for detecting and tracking opinions in on-line discussion. In *Proceedings SIGIR 2001 Workshop on Operational Text Classification*, 2001.

P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings ACL 2002*, 2002.

Y. Wang & I. H. Witten. Pace regression. Technical Report 99/12, Department of Computer Science, University of Waikato, September 1999.

I. H. Witten & E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2nd edition, 2005.