Estimating Temporal Dynamics of Human Emotions

Seungyeon Kim[*] and **Joonseok Lee**[*] and **Guy Lebanon**[†] and **Haesun Park**[*]
[*]College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
[†]Amazon, Seattle, WA, USA
{seungyeon.kim, jlee716}@gatech.edu, glebanon@gmail.com, hpark@cc.gatech.edu

Abstract

Sentiment analysis predicts a one-dimensional quantity describing the positive or negative emotion of an author. Mood analysis extends the one-dimensional sentiment response to a multi-dimensional quantity, describing a diverse set of human emotions. In this paper, we extend sentiment and mood analysis temporally and model emotions as a function of time based on temporal streams of blog posts authored by a specific author. The model is useful for constructing predictive models and discovering scientific models of human emotions.

1 Introduction

Sentiment analysis predicts the presence of a positive or negative emotion $y \in \mathbb{R}$ in a document x. Recent attempts (Mishne 2005; Kim et al. 2013) generalize sentiment analysis by introducing a multivariate response variable $y \in \mathbb{R}^d$, which corresponds to a more complex emotional state. For example, a classical model from psychology examines a two-dimensional emotional quantity in which the first dimension corresponds to sentiment and the second corresponds to the level of engagement (`aroused` vs. `calm`). Sentiment analysis and its generalizations are important tools in industry, attracting a considerable amount of attention from the research community.

The analysis of human emotions based on text has focused mostly on static analysis, that is the analysis of documents solely based on its own content ignoring temporal dependences. This paper explores a temporal model for human emotions that applies to a sequential stream of text documents written by the same author across different time points. Our model is somewhat similar to Brownian motion and the Kalman filter and generalizes the latent space emotion model in Kim et al. (2013).

Most papers on sentiment analysis use movie or item reviews, which are poorly suited to temporal modeling. Reviews relate to specific stationary truth and are unlikely to significantly change based on the time of authoring. Blog posts, however, are free expressions of the author's emotions and thus depend more on the time of authoring. Therefore, we demonstrate a temporal model on data consisting of time-stamped blog posts. We construct the sentiment or emotion ground truth from an emotion label for the text.

The temporal model is useful in two ways. First, it leads to a predictive model that estimates the current emotional state of the author within a specific time context. This predictive model is more accurate than static analysis, which ignores time information. Second, the model is useful in confirming or refuting psychological models concerning human emotions and their dependency on time. Specifically, we re-examine the circadian rhythm model from psychology and investigate its higher order generalizations and its variance across multiple individuals.

2 Related Work

Most sentiment or mood analysis work focus on extracting a richer set of features in a single document; it is somewhat orthogonal to our contribution. Pang and Lee (2008) and Liu (2012) are good surveys of this field. Mishne (2005), Généreux and Evans (2006), and Keshtkar and Inkpen (2009) are also notable as they extended binary sentiment to multivariate emotions using multiclass classifiers. Kim et al. (2013) is the most closely aligned to our work because of its similar multi-dimensional latent variable formulation.

Several studies have devoted to the temporal variation of sentiment even though they have not examined inter-document dependencies. Mishne and Maarten (2006) explored a trend of emotions, yet they did not include temporal dependencies in their model. Mao and Lebanon (2007; 2009) introduced temporal variation in topics and sentiments within a single document.

Supervised and temporal topic models are also similar to our model. Supervised topic models (Blei and McAuliffe 2007; Ramage et al. 2009) include similar graphical models except for their use of discrete latent variables and lack of temporal dependencies. Temporal topic models (Blei and Lafferty 2006; Wei, Sun, and Wang 2007; Wang, Blei, and Heckerman 2009; Hong et al. 2011), by contrast, do not use supervised label information. Conditional Random Field (CRF) (Lafferty, Pereira, and McCallum 2001) includes both dependencies; however, it does not use specific time information. Additional examples of temporal modeling of text documents include Lebanon, Mao, and Dillon (2007); Mao, Dillon, and Lebanon (2007); Lebanon and Zhao (2008).

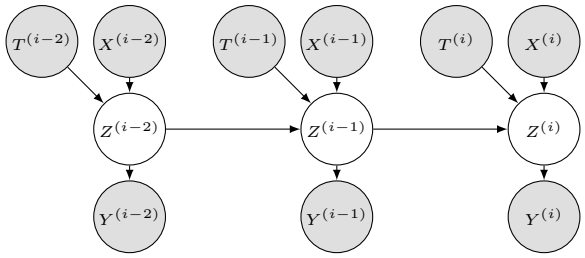

Figure 1: Graphical model of the temporal sentiment analysis model. X denote sequence of documents, T is the corresponding authoring time, Y is for the sentiment, and Z is the continuous latent variable. See text for details.

The psychological community has not employed such graphical models; however, it has devoted a great deal of effort to explore human emotions. Note that *emotion*, *affect*, and *mood* have distinguishable meanings in psychology, but we use them here interchangeably. Watson and Tellegen (1985) introduced the dimensional structures of emotions. Lewis and Granic (2002) and Verduyn et al. (2009) described the temporal dynamics of moods. Murray, Allen, and Trinder (2002) and Golder and Macy (2011) explored periodic phenomena in temporal progression of moods. Papers in the psychology literature build models mostly based on data from a limited set of human surveys. Our approach is to leverage the internet by analyzing user-generated content such as blog posts.

Our work differs from previous studies in primarily three ways. First, unlike sentiment or mood analysis work, we employ temporal dependencies between documents. Second, our work assumes continuous embedding while other supervised topic models assume a discrete set of labels (multiclass). Third, we use specific time information rather than only ordering between documents, which is covered in CRF.

3 Temporal Dynamics of Binary Sentiment

We augment the standard dataset in sentiment analysis $\{(X^{(i)}, Y^{(i)}), i = 1, \ldots, n\}$ with time stamps $T^{(i)} \in \mathbb{R}$, representing the time document $X^{(i)}$ was authored. We assume that the documents are represented as feature vectors $X^{(i)}$. The feature vector can have any document-level features such as bag-of-words or even sophisticated autoencoder features. The response variables $Y^{(i)} \in \{-1, +1\}$ are binary sentiment polarity values. We additionally assume a latent variable $Z^{(i)} \in \mathbb{R}$ associated with $X^{(i)}$ and $Y^{(i)}$ representing a continuous sentiment concept.

The introduction of the continuous latent variable serves several roles: (i) it is easier to construct temporal models in continuous state space, and (ii) the framework conveniently generalizes to mood analysis where there are a large number of emotions embedded in a low dimensional continuous space (we explore this generalization in the next section).

We assume that $Z|Y$ follows a Gaussian distribution in $\mathbb{R}$, implying that Y is a (stochastic) discretization of Z. We also assume that $Z|X$ follows a linear regression model and that $Z^{(1)}, \ldots, Z^{(n)}$ follow a Markov chain with Gaussian conditional distributions $Z^{(i)}|Z^{(i-1)}$. The formal definition appears below and the graphical model appears in Figure 1.

1. $X^{(i)} \to Z^{(i)} \to Y^{(i)}$ forms a Markov chain ($Y^{(i)}$ is independent of $X^{(i)}$ given $Z^{(i)}$).
2. The distribution of $Z|Y$ is a Gaussian with appropriate mean and variance:
$$\{Z^{(i)}|Y^{(i)} = y\} \sim \mathcal{N}(y, \sigma_y^2),\ y \in \{-1, +1\}.$$
3. The distribution of $Z|X$ follows a linear regression model (assuming the document X is represented as a vector)
$$\{Z^{(i)}|X^{(i)} = x\} \sim \mathcal{N}(\theta^\top x, \epsilon^2),\ X^{(i)}, \theta \in \mathbb{R}^k. \quad (1)$$
4. The latent variables follow a Markov chain with Gaussian conditionals. ($\Delta T = T^{(i)} - T^{(i-1)}$)
$$\{Z^{(i)}|Z^{(i-1)}\} \sim \mathcal{N}(Z^{(i-1)}, \beta\Delta T) \quad (2)$$

Assumptions 3 and 4 above can be combined to produce

$$\{Z^{(i)}|X^{(i)}, Z^{(i-1)}\} \sim \mathcal{N}(\mu^{(i)}, \sigma^{(i)})$$
$$\text{where } \sigma^{(i)} = \left((\beta\Delta T)^{-1} + \epsilon^{-2}\right)^{-1}$$
$$\mu^{(i)} = \sigma^{(i)}\left((\beta\Delta T)^{-1} Z^{(i-1)} + \epsilon^{-2}\theta^\top X^{(i)}\right).$$

For simplicity, we consider (above and in the sequel) the time points $T^{(i)}$ to be non-random, and we therefore omit them in the probability notations for example $P(Z^{(i)}|X^{(i)})$ rather than $P(Z^{(i)}|X^{(i)}, T^{(i)})$. This is analogous to fixed design in regression analysis, as opposed to random design.

We emphasize the following characteristics. First, low-dimensional latent variable $Z^{(i)}$ solely determines polarity $Y^{(i)}$ ($Y^{(i)}$ is independent of high-dimensional $X^{(i)}$ given $Z^{(i)}$). Second, $Z^{(i)}$ has a distribution that centered at $Z^{(i-1)}$ with a variance that increases with ΔT, which is in agreement with psychological observations as well as standard models in the time series literature. Third, as we do not specify $p(X)$, our model is a discriminative model. It is similar to standard discriminative structured classifiers (such as CRF) with an additional constraint for inter-document dependency by β and ΔT. It matches our intuition as temporal proximity tends to imply proximity in sentiment (for blog posts written by the same author).

3.1 Learning and Using the Model

Parameters $\eta = (\theta, \beta, \mu_y, \sigma_y)$ can be estimated by maximizing the conditional likelihood, $p(Y|X)$, of observed data.

We consider two alternatives to handle documents written by different authors: (i) estimating a single set of parameters for all authors, and (ii) estimating separate parameters for each author. In the first approach, the model is universal and can capture generic trends. The second approach fits specialized models for each author. While the first approach appears more limited than the second, it is particularly useful when some authors do not have sufficient labeled data. In either case, we maximize the likelihood function for the observed data, which integrates over the latent variables.

In the first case above, we estimate the parameter by maximizing the total sum of conditional log-likelihood (temporal

dependencies only holds for each author). Denoting the set of authors by A and an individual author as $a \in A$, we have

$$\eta = \arg\max_{\eta} \sum_{a \in A} \ell(\eta, a) \tag{3}$$

$$\begin{aligned}\ell(\eta, a) &= \log p_\eta(y_a^{(1)}, \ldots, y_a^{(n)} | x_a^{(1)}, \ldots, x_a^{(n)}) \\ &= \log \int_z p_\theta(z^{(1)}|x_a^{(1)}) \cdot \prod_{i=2}^{n} p_{\theta,\beta}(z^{(i)}|z^{(i-1)}, x_a^{(i)}) \\ &\quad \cdot \prod_{i=1}^{n} p_{\mu_y,\sigma_y}(y_a^{(i)}|z^{(i)}) dz, \end{aligned} \tag{4}$$

where $x_a^{(i)}$ and $y_a^{(i)}$ denote documents and labels associated with a specific author $a \in A$ and the integral over z represents integration over the latent variables $z^{(i)}$. In the second case, the log-likelihood is the same as above except that we have multiple terms of log-likelihood for each author with different parameters $\eta_a, a \in A$. It can possibly describe behaviors of an author in depth (some authors may have stronger temporal dependency or lesser).

Inference To predict the most likely polarity of a given temporal document (in test time), we compute below the most likely value. Note that we use the previous time stamped documents to improve the estimation accuracy.

$$\hat{y}^{(i)} = \arg\max_{y^{(i)}} p(y^{(i)}|x^{(i)}, \ldots, x^{(1)}) \tag{5}$$

$$\begin{aligned} &= \arg\max_{y^{(i)}} \int \cdots \int p(y^{(i)}|z^{(i)}) \\ &\quad \cdot \prod_{j=1}^{i} p(z^{(j)}|x^{(j)}, z^{(j-1)}) dz^{(1)} \ldots dz^{(i)}. \end{aligned} \tag{6}$$

Approximation and Implementation There are several approaches for computing (4) including numeric integration, Laplace approximation, and Markov Chain Monte Carlo (MCMC). We use a simpler approximation that replaces the Gaussian distribution over Z with a Dirac's delta centered at the most likely value of z. This approximation replaces the integral with the integrand, evaluated at the most likely value of the latent variable. Naturally, the approximation quality increases as the variance of the Gaussian decreases.

$$\int \mathcal{N}(z\,; z^*, \sigma) g(z) dz \approx c(\sigma) \int \delta(z - z^*) g(z) dz = c(\sigma) g(z^*).$$

Applying this approximation to (4) we get:

$$\begin{aligned}\ell &\approx \log p(y^{(1)}, \ldots, y^{(n)}, z^{(1)} = z^{*(1)}, \ldots | x^{(1)}, \ldots, x^{(n)}) \\ &= \sum_{i=1}^{n} \log p_\theta(z^{*(1)}|x^{(1)}) + \sum_{i=2}^{n} \log p_{\theta,\beta}(z^{*(i)}|z^{*(i-1)}, x^{(i)}) \\ &\quad + \sum_{i=1}^{n} \log p_{\mu_y,\sigma_y}(y^{(i)}|z^{*(i)}) \end{aligned} \tag{7}$$

where $z^{*(i)} = \arg\max_z p(z^{*(i)}|z^{*(i-1)}, x^{(i)}) p(y^{(i)}|z^{*(i)})$

$$= \left({\sigma^{(i)}}^{-1} + {\sigma_y}^{-1}\right)^{-1} \left({\sigma^{(i)}}^{-1}\mu^{(i)} + {\sigma_y}^{-1}\mu_y\right).$$

$\pm$	**Assignments of moods**
+1	excited, pleased, good, cheerful, amused, hopeful, bouncy, chipper, thoughtful, accomplished
-1	cold, exhausted, sleepy, tired, bored, sick, sore, uncomfortable, depressed, sad, annoyed

Table 1: Polarity assignments of moods

The maximum likelihood estimator for θ, $\hat{\theta} = \arg\max_\theta$

$$\left[\log p(z^{*(1)}|x^{(1)}) + \sum_{i=2}^{n} \log p(z^{*(i)}|z^{*(i-1)}, x^{(i)})\right]$$

is conveniently given in closed form. Since the conditional probabilities are Gaussian, the MLE is equivalent to a least squares linear regression model $\theta^\top X = W$ where $X = [x^{(1)}, \ldots, x^{(n)}]$ and $W \in \mathbb{R}^n$ is

$$W = \begin{bmatrix} z^{*(1)} \\ \vdots \\ \epsilon^2\left[\left((\beta\Delta T)^{-1} + \epsilon^{-2}\right) z^{*(i)} - (\beta\Delta T)^{-1} z^{*(i-1)}\right] \\ \vdots \end{bmatrix}.$$

After obtaining the MLE $\hat{\theta}$, we can fix that parameter value and compute the MLE for the remaining parameters $(\hat{\beta}, Z^*, \hat{\mu}_y, \hat{\sigma}_y)$ using standard gradient based optimization. We then re-calculate $\hat{\theta}$ and iterate until convergence.

The approximation above can also be used in test time prediction. It proceeds as above by replacing the integrals over the latent variables Z by the integrand evaluated at the most likely value of the latent variables.

3.2 Experiment

Dataset Most standard sentiment analysis datasets (Pang, Lee, and Vaithyanathan 2002; Wiebe, Wilson, and Cardie 2005; Ganu, Elhadad, and Marian 2009) focus on a sentiment concept corresponding to opinions or reviews on specific topics. This sentiment concept is unlikely to vary significantly with time as much because it reflect the author's opinion about a specific issue (however, see Koren (2009) that discovers some temporal effects in movie reviews). We choose instead to model blog posts, which depend more significantly on time.

We gathered data by crawling a popular blog service, Livejournal[1] from May 2010 to July 2011. Livejournal provides time-stamps as well as emotion annotation that reflect the author's mood of a single blog post (one annotation per one blog post). The authors are offered the use of a wide range of emoticons and also offers a free text emotion annotation. The crawling resulted in two million documents authored by 315K authors, and about 20% of the authors annotated their posts with such annotation.

Since most authors do not provide more than 2 documents (the median number of blog post by an author), we selected the top 50 most frequently publishing authors. We also had to remove spam authors who kept posting the same content

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

Methods	F1	Accuracy
temporal sentiment method	**0.7596**	**0.8058**
temporal linear-chain CRF	0.7130	0.7742
temporal VARX(1)	0.6554	0.7172
non-temporal logistic regression	0.7109	0.8016
non-temporal SVM	0.6555	0.7557
non-temporal SLDA	0.5093	0.7379
non-temporal naive Bayes	0.6915	0.7373

Table 2: Test set F1 and accuracy results for predicting sentiment polarity. Bold face shows statistically significant improvement over other competitors (t-test, 95% confidence).

repeatedly with random emotions. We finally obtained 19 authors with 64 different emotions after removing rare emotions that appeared in less than 20 documents.

We want to note that most authors don't write frequently, which makes estimating their emotion harder. For example, the 50th most active author had only 52 documents for training. It is a common case when we handle social media data. We expect our method that exploits temporal dependencies to lead to a more accurate model than non-temporal models in this sparse setting.

In this section, we converted the emotions to polarity labels $\{-1, +1\}$, while ignoring neutral emotions as indicated by Table 1. This procedure yielded 6,295 documents with 0.27:0.73 (negative:positive) class distributions. Section 4.1 describes experiments on the entire set of emotions including the neutral emotions.

For document-level features, $X^{(i)}$, we selected negation-handled term frequencies. Specifically, we tokenized and stemmed (Krovetz) the documents, and added new features for negated words as is commonly done in sentiment analysis (Kennedy and Inkpen 2006). For example, this implies that negation words like "not" or "isn't" followed by "good" will yield negated tokens such as "not-good." This procedure yielded 43,910 features. To create the final document representation, we used the square root of the normalized term frequency of the tokens since (i) it is closely associated with the Hellinger distance and the multinomial Fisher geometry that have nice theoretical properties (Čencov 1982), and (ii) it often leads to improved modeling accuracy (see Lafferty and Lebanon (2005), Lebanon (2005a; 2005c; 2005b), and Dillon et al. (2007) for examples of using Hellinger distances in text modeling and interpreting it in terms of information geometry).

It is possible other sophisticated features such as dependency tree features or auto encoder features can be used for $X^{(i)}$; however, the term frequency feature is sufficient for demonstrating the model and contrasting it with standard sentiment analysis models. Our main contribution is presenting the manifold Z which is applicable regardless of the document-level feature extraction.

A few statistics concerning the dataset are: (i) the average document length is 71 words (± 89), (ii) most documents (55%) have less than 50 words, (iii) there are a few very long documents (0.01%) with more than 500 words, and (iv) the average word length is 8.33 characters. This denotes our blog posts are relatively sparser than other popular sentiment dataset such as the movie review dataset (Pang, Lee, and Vaithyanathan 2002).

Methods	Time (sec)
temporal sentiment method	16.96
temporal linear-chain CRF	3352.81
temporal VARX(1)	9.17
non-temporal logistic regression	48.39
non-temporal SVM	27.26
non-temporal SLDA	1047.59
non-temporal naive Bayes	1.07

Table 3: Training time comparison of each method

Classification One of the primary tasks in sentiment analysis is predicting the sentiment polarity $Y \in \{-1, +1\}$ of a document X. Table 2 compares seven different methods in sentiment prediction with our model using shared parameters across all authors (Equation 3). With larger dataset, we may use author-specific parameters to improve the accuracy.

We compare our temporal sentiment method with non-temporal classification baselines including a well known supervised topic model (Supervised Latent Dirichlet Allocation[2] (Blei and McAuliffe 2007)), SVM[3], logistic regression with L_2 regularization, and naive Bayes classifier. We also compare our model with temporal models such as Conditional Random Field[4] (Lafferty, Pereira, and McCallum 2001) and VARX(1)[5], $y^{(t)} = a + X^{(t)}\beta + Ay^{(t-1)}$, which is the most relevant vector autoregressive model. The inputs of VARX(1) model are given $\{-1, +1\}$ accordingly and the output is interpreted as a binary classification by its polarity.

Training and test set were split 50:50 by post-by-post random (not author-based). Post orderings of each author were recovered using timestamps $T^{(i)}$ and author ID after the split. All regularization parameters were chosen by a validation set on training examples using grid search of 5 candidates. We tried various sizes of latent topics on SLDA (2 to 20) and included the best result.

Our model demonstrates statistically significant improvement in both F1 measure and classification accuracy (t-test with 100 random trials) compared to all non-temporal methods and other temporal methods. SLDA performed relatively poorly especially on F1, perhaps due to the extreme sparsity. CRF and VARX(1) do not explicitly model the time gap (ΔT) between observations and rather only consider the ordering of the time stamps unlike the proposed model. This explains the lower performance of CRF, which is generally considered a top performing model.

Table 3 shows average training time of each methods. Our method (temporal sentiment) takes much shorter time

[2]Gibbs sampling implementation: `https://github.com/michaelchughes/SuperTopicModels`

[3]LibSVM (ν-SVM, RBF kernel): `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

[4]Linear-chain CRF from Kevin Murphy: `http://www.cs.ubc.ca/~murphyk/Software/CRF/crf.html`

[5]Matlab implementation

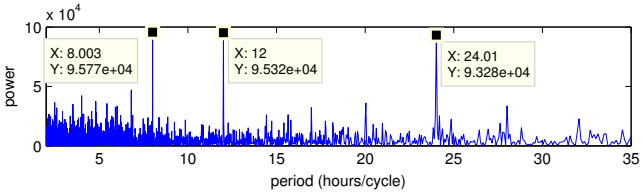

Figure 2: Fourier components of the latent variable in the global model. There are three significant periodic components representing the periods: 8 hours, 12 hours, and 24 hours (circadian rhythm).

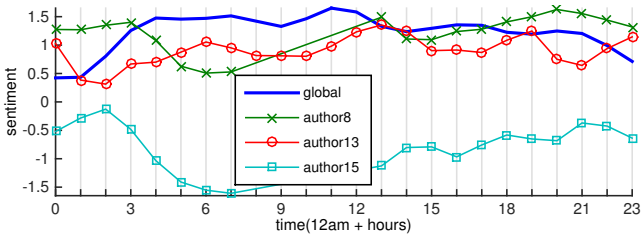

Figure 3: Hourly pattern of global sentiment and selected authors. The y axis correspond to the latent variable of the model (higher values correspond to stronger positive sentiment). See text for details.

to train the model compared to methods that have comparable performance (such as CRF, SVM or logistic regression). In fact, it is far faster than CRF and SLDA, which are considered to be one of the state-of-the-art methods.

Periodicity The temporal model can also be used for analyzing temporal variations in the emotions of authors. Figure 2 investigates the periodic behavior of the temporal sentiment by displaying the Fourier components of the latent variable. The x axis of the graph measures period (one over frequency) rather than traditional Fourier frequency for better interpretation in our context.

The peaks in graph show significant periodic components at 8 hours, 12 hours, and 24 hours. Interestingly, the strong 24 hours periodicity matches the circadian rhythm discovery from chronobiology and psychology research (Murray, Allen, and Trinder 2002; Golder and Macy 2011). Specifically, this confirms the work of (Murray, Allen, and Trinder 2002) that explored a circadian component in positive affect. The confirmation is noteworthy as our model was constructed from blog posts, while they surveyed human subjects in a controlled environment.

Hourly Pattern From Figure 2, we found 24 hours of periodicity. We investigate this finding further by visualizing hourly average sentiment values (as measured by the latent variable) for models trained on data of specific authors and for the globally-trained model (Figure 3).

The figure shows interesting observations that agree our intuition. First, the trend of global model is well aligned with generic daily schedules. Four local maxima at 6am, 11am, 16pm and 20pm match to positive daily events: wake-up time, lunch break, office closing time, and dinner time. There are notable minima in late night (0am-3am) and the end of lunch break (1pm). The highest and lowest sentiment values are achieved around noon and midnight, respectively. Second, some authors have different temporal patterns, which may indicate different time zones or different life styles. For example, author 8 and 15 exhibit low sentiment in the early morning and high sentiment late at night. Third, the hourly trend shows characteristics of an author. A sentiment structure of author 15 shows an overall lower sentiment compared to others. When we manually visit the actual blog site of the author, we observed many negative annotations. These observations can be useful in psychological studies as well as marketing and advertisement sciences.

4 Temporal Dynamics of Multivariate Emotions

We now extend the temporal dynamics model to the case where a richer concept describing a diverse set of emotions rather than having a one dimensional polarity. Such emotions, for example `happy`, `sad`, `excited`, and `tired` are correlated with the polarity, but they offer an opportunity to construct a more fine grained model of the author's emotion.

Since there is a relatively large set of possible emotions, and these emotions are related to one another, we avoid constructing a separate temporal dynamics model for each individual binary emotion. Instead, we extend the formalism of Kim et al. (2013), where a large set of emotions are embedded in a low dimensional Euclidean space. We thus generalize the model of the previous section by increasing the dimensionality of the latent variable Z. We make use of the same notation as Section 3, but now $Y^{(i)} \in \{1, 2, \ldots, |C|\}$, and the latent variable is a vector $Z^{(i)} \in \mathbb{R}^l$.

The largest difference between our work with Kim et al. (2013) is that their work lacks temporal dependence. They assume blog posts are independent from each other while our model makes better use of data considering previous observations. Another difference is that we consider both individual-level and global-level emotion manifolds, as described in the previous section.

The assumptions in Section 3 are now extended to adapting multivariate emotions. Assumption 2 now has a multivariate Gaussian instead of the univariate Gaussian: $\{Z^{(i)}|Y^{(i)} = y\} \sim \mathcal{N}(\mu_y, \Sigma_y)$; the centroids ($\mu_y$) correspond to $\{-1, +1\}$ from the previous section that described positive and negative polarity in the latent space. Assumption 3 is extended to multi-response regression: $\{Z^{(i)}|X^{(i)} = x\} \sim \mathcal{N}(\Theta x, \epsilon^2 \mathsf{I}), \Theta \in \mathbb{R}^{l \times k}$, and a spherical covariance is used instead of a scalar in Assumption 4: $\{Z^{(i)}|Z^{(i-1)}\} \sim \mathcal{N}(Z^{(i-1)}, \beta \cdot \Delta T \cdot \mathsf{I})$.

We additionally introduce the fifth and last assumption similar to that in Kim et al. (2013).

5. $\forall y \in C$, distances between $\{\mathsf{E}[Z^{(i)}|Y^{(i)} = y]\}$ are similar to the corresponding distances in $\{\mathsf{E}[X^{(i)}|Y^{(i)} = y]\}$.

This assumption enables us to estimate μ_y, Σ_y by multi-dimensional scaling (MDS) on empirical averages corresponding to $\mathsf{E}[X|Y = y]$.

It is worth mentioning our model incorporates two different types of proximities: temporal proximities between $Z^{(i)}$ and $Z^{(i-1)}$ and spatial proximities between $\mathsf{E}[Z^{(i)}|Y^{(i)}]$ and

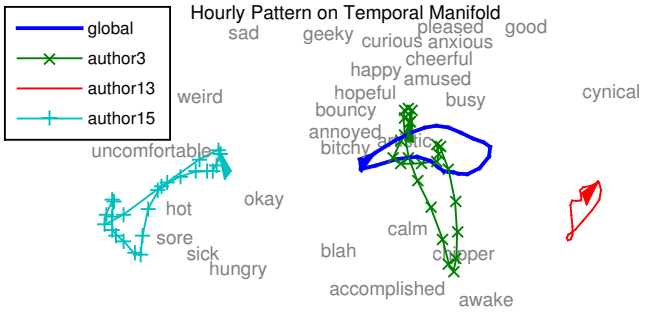

Figure 4: Hourly trends of global model and selected authors on the first two dimensions of the manifold (smoothed). Gray words show $\mathsf{E}[Z|Y = y]$. The arrow shows the start of the day (12am) and direction of the progression of each circle. There is clear separation between day and night time.

$\mathsf{E}[X^{(i)}|Y^{(i)}]$. We also expect the dimensionality l of the latent space $Z^{(i)}$ to be smaller than the number of emotions $l \ll |C|$ and much smaller than the dimensionality of $X^{(i)}$. This means that $Z^{(i)}$ serves as a latent low-dimensional variable that connects two observed high dimensional variables.

To estimate the model, we start by MDS to estimate μ_y, Σ_y. The dimension of latent space Z is bounded by $|C| - 1$ consequently. We then follow the maximum likelihood procedure and approximation as described in the previous section.

4.1 Experiments

Dataset We now consider full emotion labels (including all neutral emotions) from Section 3.2. The data contains 11,659 documents with 43,910 features and 64 emotions. The label frequencies varied between 0.0018 (21 documents) to 0.1441 (1680 documents), with an average label frequency of 0.0156 (182 documents).

Hourly Pattern Figure 4 shows temporal trajectories of the latent variable in several models based on individuals, as well as a global model on the first two dimensions of the model. The first two dimensions seem to capture sentiment level (horizontally) and energy level (vertically). Gray words show centroids of corresponding emotion labels.

The global model shows a progression from negative sentiment (left side) to positive sentiment (right side) from 12am, and the progression is reversed later on 12pm similar to Figure 3. Each author shows their unique progression style and location. For example, author 15 is exceptional in its location (close to negative emotions) and its progression (counter-clockwise) as opposed to the frequently observed clockwise progression; this means the author has emotional transitions in an opposite order to most people.

Classifying Emotions We now consider the task of predicting the emotion of a given text document. Compared to classifying a binary sentiment polarity, predicting the multiclass emotion concept is difficult due to the large number of interrelated emotions, some of which having only a small set of labeled documents.

Table 4 shows the test set classification performance of predicting the emotion (out of a set of $|C| = 64$ possible emotions). We compared our temporal model with global parameter setting, the non-temporal model (Kim et al. 2013), SLDA (Blei and McAuliffe 2007), logistic regression, and naive Bayes. CRF failed to converge in weeks of running time. Details of experimental configuration remain the same as in Section 3.2 except that (i) the size of latent topics on SLDA was varied from $2{\cdot}|C|$ to $5{\cdot}|C|$ and (ii) the experiment was repeated 50 times by random split.

	Macro F1	Accuracy
temporal model (proposed)	**0.2552**	0.4381
non-temporal model	0.2522	0.4370
logistic regression	0.1618	0.4329
SLDA	0.1131	0.3358
naive Bayes	0.0103	0.1405

Table 4: Test set F1 and accuracy results for predicting emotion among 64 emotions. Bold face shows statistically significant improvement over other competitors (t-test, 95% confidence). See text for details.

The improvement in classification accuracy is noticeable, but not statistically significant; however, the improvement in macro F1 is statistically significant. This is notable as our class distribution is highly unbalanced and F1 is the measure that is often regarded as more informative than accuracy in the multiclass case with unbalanced class frequencies.

Above, the dimension of the latent space l is $|C|-1$. However, reducing the dimensionality to $0.7|C|$, $0.45|C|$, and $0.19|C|$ reduces the accuracy of the full model to 90%, 80%, and 70% respectively. This shows that the manifold includes most of its information in only a few dimensions. This classification experiment includes tiny classes that have a few documents (minimum 20 documents). When we performed the same experiment while excluding tiny classes (minimum 100 documents), the non-temporal method (Kim et al. 2013) performed better in terms of emotion classification accuracy. We conclude that our model is well suited to handling small classes, which is often the case in self reported mood data.

5 Summary

We presented a temporal statistical model for modeling binary sentiment polarity values and multivariate emotions. Our model uses temporally-dependent continuous latent variables in order to capture relationships between various emotions and observations. We examined a wide variety of applications, including sentiment or emotion prediction of time-stamped documents and scientific research into the temporal variations of human emotions. Our experimental results demonstrate improved statistical modeling and confirmation of discoveries from the psychology literature concerning the static and temporal structure of human emotions.

Acknowledgments

This work was supported in part by NSF Grant CCF-1348152 and DARPA XDATA Grant FA8750-12-2-0309. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Blei, D., and Lafferty, J. 2006. Dynamic topic models. In *Proc. of the International Conference on Machine Learning*.

Blei, D., and McAuliffe, J. 2007. Supervised topic models. *Advances in Neural Information Processing Systems*.

Čencov, N. N. 1982. *Statistical Decision Rules and Optimal Inference*. American Mathematical Society.

Dillon, J.; Mao, Y.; Lebanon, G.; and Zhang, J. 2007. Statistical translation, heat kernels, and expected distances. In *Uncertainty in Artificial Intelligence*, 93–100. AUAI Press.

Ganu, G.; Elhadad, N.; and Marian, A. 2009. Beyond the stars: Improving rating predictions using review text content. In *International Workshop on the Web and Databases*.

Généreux, M., and Evans, R. 2006. Distinguishing affective states in weblog posts. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*.

Golder, S., and Macy, M. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333(6051):1878–1881.

Hong, L.; Yin, D.; Guo, J.; and Davison, B. 2011. Tracking trends: incorporating term volume into temporal topic models. In *Proc. of International Conference on Knowledge Discovery and Data mining*.

Kennedy, A., and Inkpen, D. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence* 22(2):110–125.

Keshtkar, F., and Inkpen, D. 2009. Using sentiment orientation features for mood classification in blogs. In *IEEE International Conference on Natural Language Processing and Knowledge Engineering*.

Kim, S.; Li, F.; Lebanon, G.; and Essa, I. 2013. Beyond sentiment: The manifold of human emotions. In *Proc. of the International Conference on Artificial Intelligence and Statistics*.

Koren, Y. 2009. Collaborative filtering with temporal dynamics. In *Proc. of International Conference on Knowledge Discovery and Data mining*.

Lafferty, J., and Lebanon, G. 2005. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research* 6:129–163.

Lafferty, J.; Pereira, F.; and McCallum, A. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proc. of the International Conference on Machine Learning*.

Lebanon, G., and Zhao, Y. 2008. Local likelihood modeling of the concept drift phenomenon. In *Proc. of the 25th Intrenational Conference of Machine Learning*.

Lebanon, G.; Mao, Y.; and Dillon, J. 2007. The locally weighted bag of words framework for documents. *Journal of Machine Learning Research* 8:2405–2441.

Lebanon, G. 2005a. Axiomatic geometry of conditional models. *IEEE Transactions on Information Theory* 51(4):1283–1294.

Lebanon, G. 2005b. Information geometry, the embedding principle, and document classification. In *Proc. of the 2nd International Symposium on Information Geometry and its Applications*, 101–108.

Lebanon, G. 2005c. *Riemannian Geometry and Statistical Machine Learning*. Ph.D. Dissertation, Carnegie Mellon University, Technical Report CMU-LTI-05-189.

Lewis, M. D., and Granic, I. 2002. *Emotion, development, and self-organization: Dynamic systems approaches to emotional development*. Cambridge University Press.

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* 5(1).

Mao, Y., and Lebanon, G. 2007. Isotonic conditional random fields and local sentiment flow. In *Advances in Neural Information Processing Systems 19*, 961–968.

Mao, Y., and Lebanon, G. 2009. Generalized isotonic conditional random fields. *Machine Learning* 77(2-3):225–248.

Mao, Y.; Dillon, J.; and Lebanon, G. 2007. Sequential document visualization. *IEEE Transactions on Visualization and Computer Graphics* 13(6):1208–1215.

Mishne, G., and Maarten, R. 2006. Capturing global mood levels using blog posts. In *AAAI Symposium on Computational Approaches to Analysing Weblogs*.

Mishne, G. 2005. Experiments with mood classification in blog posts. In *Workshop on Stylistic Analysis Of Text For Information Access*.

Murray, G.; Allen, N.; and Trinder, J. 2002. Mood and the circadian system: Investigation of a circadian component in positive affect. *Chronobiology international* 19(6):1151–1169.

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* 2:1–135.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*.

Verduyn, P.; Delvaux, E.; Coillie, H. V.; Tuerlinckx, F.; and Mechelen, I. V. 2009. Predicting the duration of emotional experience: Two experience sampling studies. *Emotion* 9(1):83.

Wang, C.; Blei, D.; and Heckerman, D. 2009. Continuous time dynamic topic models. In *Proc. of Uncertainty in Artificial Intelligence*.

Watson, D., and Tellegen, A. 1985. Toward a consensual structure of mood. *Psychological bulletin* 98(2):219–235.

Wei, X.; Sun, J.; and Wang, X. 2007. Dynamic mixture models for multiple time-series. In *International Joint Conferences on Artificial Intelligence*.

Wiebe, J.; Wilson, T.; and Cardie, C. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39(2):165–210.