Beyond Sentiment: The Manifold of Human Emotions

Seungyeon Kim, Fuxin Li, Guy Lebanon, and Irfan Essa
College of Computing
Georgia Institute of Technology
{seungyeon.kim@, fli@cc., lebanon@cc., irfan@cc.}gatech.edu

Abstract

Sentiment analysis predicts the presence of positive or negative emotions in a text document. In this paper we consider higher dimensional extensions of the sentiment concept, which represent a richer set of human emotions. Our approach goes beyond previous work in that our model contains a continuous manifold rather than a finite set of human emotions. We investigate the resulting model, compare it to psychological observations, and explore its predictive capabilities. Besides obtaining significant improvements over a baseline without manifold, we are also able to visualize different notions of positive sentiment in different domains.

1 Introduction

Sentiment analysis predicts the presence of a positive or negative emotion y in a text document x. Despite its successes in industry, sentiment analysis is limited as it flattens the structure of human emotions into a single dimension. "Negative" emotions such as `depressed`, `sad`, and `worried` are mapped to the negative part of the real line. "Positive" emotions such as `happy`, `excited`, and `hopeful` are mapped to the positive part of the real line. Other emotions like `curious`, `thoughtful`, and `tired` are mapped to scalars near 0 or are otherwise ignored. The resulting one dimensional line loses much of the complex structure of human emotions. Note that *emotion*, *affect*, and *mood* have distinguishable meanings in psychology, but we use them here interchangeably.

An alternative that has attracted a few researchers in recent years is to construct a finite collection of emotions and fit a predictive model for each emotion $\{p(y_i|x), i = 1, \dots, C\}$. A multi-label variation that allows a document to reflect more than a single emotion uses a single model $p(y|x)$ where $y \in \{0,1\}^C$ is a binary vector corresponding to the presence or absence of emotions. In contrast to sentiment analysis, this approach models the higher order structure of human emotions.

There are several significant difficulties with the above approach. First, it is hard to capture a complex statistical relationship between a large number of binary variables (representing emotions) and a high dimensional vector (representing the document). It is also hard to imagine a reliable procedure for compiling a finite list of all possible human emotions. Finally, it is not clear how to use documents expressing a certain emotion, for example `tired`, in fitting a model for predicting a similar one, for example `sleepy`. Using labeled documents only in fitting models predicting their denoted labels ignores the relationship among emotions, and is problematic for emotions with only a few annotated.

We propose an alternative approach that models a stochastic relationship between the document X, an emotion label Y (such as `sleepy` or `happy`), and a position on the mood manifold Z. We assume that all the emotional aspects in the documents are captured by the manifold, implying that the emotion label Y can be inferred directly from the projection Z of the document on the manifold, without needing to consult the document again.

The key assumption in constructing the manifold Z is that the spatial relationship between $X|Y = j, j = 1, \dots, C$ is similar to the spatial relationship between $Z|Y = j, j = 1, \dots, C$ (see assumption 4 in the next section).

2 Related Work

Studying emotions or affects and their relations is one of the major goals of the psychology community. There are two main approaches: categorical or dimensional.

Appearing in Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31. Copyright 2013 by the authors.

Our focus is on dimensional analysis, as described in (Russell, 1979, 1980; Shaver et al., 1987; Watson and Tellegen, 1985; Watson et al., 1988; Tellegen et al., 1999; Larsen and Diener, 1992; Yik et al., 2011).

Our work deviates from research in psychology in that we construct our model based on a large collection of annotated documents rather than an experiment with a small number of human subjects. In addition, our model has much higher dimensionality compared to traditional 2-3 dimensions used in psychology.

Sentiment analysis is a significant research direction within the natural language processing community. Pang and Lee (2008) is a recent survey of research in this area. Some recent methods are (Nakagawa et al., 2010; Socher et al., 2011).

Alm (2008) summarizes affect analysis in text and speech, while Holzman and Pottenger (2003) uses linguistic features to detect emotions in internet chatting. The work described in (Rubin et al., 2004; Strapparava and Mihalcea, 2008) classified data using a categorical model suggested by psychological literature. Mishne (2005) and Généreux and Evans (2006) examine a similar analysis task using blog posts with standard machine learning techniques, while Keshtkar and Inkpen (2009) exploit a mood hierarchy to improve classification results. The work described in (Strapparava and Valitutti, 2004; Quan and Ren, 2009; Mohammad and Turney, 2011) address the task of constructing a useful corpus for emotion analysis.

Previous work handles the mood prediction problem as multiclass classification with discrete labels. Our work stands out in that it assumes a continuous mood manifold and thus develops an inherently different learning paradigm. Our logistic regression baseline is generally considered equivalent or better than the ones in related work using SVM (Mishne, 2005; Généreux and Evans, 2006), Naive Bayes (Strapparava and Mihalcea, 2008). Keshtkar and Inkpen (2009) exploited a user-supplied emotional hierarchy which is an additional assumption that we do not have.

3 The Statistical Model

We make the following four modeling assumptions concerning the document X, the discrete emotion label $Y \in \{1, 2, \ldots, C\}$, and the position on the continuous mood manifold $Z \in \mathbb{R}^l$.

1. We have the graphical structure: $X \to Z \to Y$, implying that the emotion label $Y \in \{1, \ldots, C\}$ is independent of the document X given Z.
2. The distribution of $Z \in \mathbb{R}^l$ given a specific emotion label $Y = y$ is Gaussian
$$\{Z|Y = y\} \sim \mathcal{N}(\mu_y, \Sigma_y). \tag{1}$$
3. The distribution of Z given the document X (typically in a bag of words or n-gram representation) is a linear regression model
$$\{Z|X = x\} \sim \mathcal{N}(\theta^\top x, \Sigma_x).$$
4. The distances between the vectors in
$$\{\mathsf{E}(Z|Y = y) : y \in C\}$$
are similar to the corresponding distances in
$$\{\mathsf{E}(X|Y = y) : y \in C\}$$

We make the following observations.

- The first assumption implies that the emotion label Y is simply a discretization of the continuous Z. It is consistent with well known research in psychology (see Section 2) and with random projection theory, which state that it is often possible to approximate high dimensional data by projecting it on a low dimensional continuous space.
- While X, Y are high dimensional and discrete, Z is low dimensional and continuous. This, together with the conditional independence in assumption (1) above, implies a higher degree of accuracy than modeling directly $X \to Y$. Intuitively, the number of parameters is on the order of $\dim(X) + \dim(Y)$ as opposed to $\dim(X)\dim(Y)$.
- The Gaussian models in assumptions 2 and 3 are simple, and lead to efficient computational procedures. We also found them to work well in our experiments. The model may be easily adapted, however, to more complex models such as mixture of Gaussians or non-linear regression models (for example, we experimented with quadratic regression models).
- Assumption 4 suggests that we can estimate $\mathsf{E}(Z|Y = y)$ for all $y \in C$ via multidimensional scaling. MDS finds low dimensional coordinates for a set of points that approximates the spatial relationship between the points in the original high dimensional space.
- The models in assumptions 2 and 3 are statistical and can be estimated from data using maximum likelihood.
- The four assumptions above are essential in the sense that if any one of them is removed, we will not be able to consistently estimate the true model.

3.1 Fitting Parameters and Using the Model

Motivated by the fourth modeling assumption, we determine the parameters $\mu_y = \mathsf{E}(Z|Y = y), y \in C$ by running multidimensional scaling (MDS) or Kernel PCA on the empirical versions of $\{\mathsf{E}(X|Y = y) : y \in C\}$, which are the class averages $\frac{1}{n_k}\sum_{y^{(i)}=k} x^{(i)}$

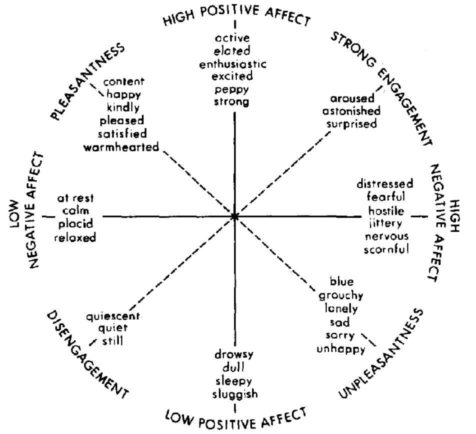

Figure 1: The two-dimensional structure of emotions from (Watson and Tellegen, 1985). We can interpret top-left to bottom-right axis as expressing sentiment polarity and the top-right to bottom-left axis as expressing engagement.

(n_k is the number of documents belonging to category k).

We estimate the parameter θ, defining the regression $X \to Z$, by maximizing the likelihood

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \sum_i \log p(y^{(i)}|x^{(i)}) \qquad (2)\\ &= \arg\max_{\theta} \sum_i \log \int_Z p(y^{(i)}|z) p_\theta(z|x^{(i)}) dz\\ &= \arg\max_{\theta} \sum_i \log \int_Z p(z|y^{(i)}) \frac{p(y^{(i)}) p_\theta(z|x^{(i)})}{\sum_y p(z|y)p(y)} dz.\end{aligned}$$

The covariance matrices Σ_y of the Gaussians $Z|Y=y$, $y = 1, \ldots, C$ may be estimated by computing the empirical variance of Z values simulated from $p_{\hat{\theta}}(Z|X^{(i)})$ for all documents $X^{(i)}$ possessing the right labels $Y^{(i)} = y$. A more computationally efficient alternative is computing the empirical variance of the most likely $\hat{Z}^{(i)}$ values corresponding to documents possessing the appropriate label $Y^{(i)} = y$:

$$\hat{Z}^{(i)} = \arg\max_z p_{\hat{\theta}}(Z = z|X^{(i)}) = \hat{\theta}^\top X^{(i)}. \qquad (3)$$

Given a new test document x, we can predict the most likely emotion with

$$\begin{aligned}\hat{y} &= \arg\max_y \int p(y, z|x) dz\\ &= \arg\max_y \int p(y|z) p_{\hat{\theta}}(z|x) dz. \qquad (4)\end{aligned}$$

But in many cases, the distribution $p(Z|X)$ provides more insightful information than the single most likely emotion Y.

3.2 Approximating High Dimensional Integrals

Some of the equations in the previous section require integrating over $Z \in \mathbb{R}^l$, a computationally difficult task when l is not very low. There are, however, several ways to approximate these integrals in a computationally efficient way.

The most well-known approximation is probably Markov chain Monte Carlo (MCMC). Another alternative is the Laplace approximation. A third alternative is based on approximating the Gaussian pdf with Dirac's delta function, also known as an impulse function, resulting in the approximation

$$\begin{aligned}\int N(z\,; \mu, \Sigma) g(z)\, dz &\approx c(\Sigma) \int \delta(z-\mu) g(z)\, dz\\ &= c(\Sigma) g(\mu). \qquad (5)\end{aligned}$$

A similar approximation can also be derived using Laplace's method. Obviously, the approximation quality increases as the variance decreases.

Applying (5) to (2) we get

$$\begin{aligned}\hat{\theta} &\approx \arg\max_{\theta} \sum_i \log \frac{p(y^{(i)}) p_\theta(z^{(i)^*}|x^{(i)})}{\sum_y p(z^{(i)^*}|y)p(y)}\\ &= \arg\max_{\theta} \sum_i \log p_\theta(z^{(i)^*}|x^{(i)}) \qquad (6)\end{aligned}$$

where $z^{(i)^*} = \arg\max_z p(z|y^{(i)}) = E(Z|y^{(i)})$, which is equivalent to a least squares regression.

Applying (5) to (4) yields a classification rule

$$\hat{y} \approx \arg\max_y p\left(y \middle| Z = \arg\max_z p_{\hat{\theta}}(z|x)\right). \qquad (7)$$

3.3 Implementation

In estimating the covariance matrices of a Gaussian $P(Z|Y=y)$, it is sometimes assumed that each class has the same covariance matrix, leading to linear discriminant analysis (LDA) as the optimal Bayes classifier. The alternative assumption that the covariance matrices for each class is different leads to quadratic discriminant analysis (QDA) as the optimal Bayes classifier.

We consider both assumptions and three different models for the covariance matrices: full covariance, diagonal covariance, and linear combination of full covariance and spherical covariance (standard regularization technique):

$$\hat{\Sigma}' = (1-\lambda)\hat{\Sigma} + \lambda \left(\sum_{i=1}^{C} \hat{\Sigma}_{ii} \right) I \qquad \text{(LDA)}$$

$$\hat{\Sigma}'_y = (1-\lambda)\hat{\Sigma}_y + \lambda \left(\sum_{i=1}^{C} [\hat{\Sigma}_y]_{ii} \right) I \qquad \text{(QDA).}$$

In either case we used a C dimensional ambient space (C equals the number of emotions) and the approximation (7).

Due to the high dimensionality of X, it may be useful to estimate $\hat{\theta}$ using ridge regression, rather than least squares regression. In this case, we update the estimate $E(Z|Y = y)$ in third stage, based on the ridge estimate $\hat{\theta}$.

One interpretation of our model $X \to Z \to Y$ is that Z forms a sufficient statistic of X for Y. We can thus consider adapting a wide variety of predictive models (for example, logistic regression or SVM) on $Z \mapsto Y$. These discriminative classifiers are trained on $\{(\hat{Z}^{(i)}, Y^{(i)}), i = 1, \ldots, n\}$.

4 Experiments

4.1 Datasets

We used crawled Livejournal[1] data as the main dataset. Livejournal is a popular blog service that offers emotion annotation capabilities to the authors. About 20% of the blog posts feature these optional annotations in the form of emoticons. The annotations may be chosen from a pre-defined list of possible emotions, or a novel emotion specified by the author. We crawled 15,910,060 documents and selected 1,346,937 documents featuring the most popular 32 emotion labels (in respect to the number of documents annotated in). It is a significantly larger dataset compare to similar works: 1,000 (Strapparava and Mihalcea, 2008), 346,723 (Généreux and Evans, 2006) and 345,014 (Mishne, 2005) documents.

We used Indri from the Lemur project[2] to extract term frequency features while tokenizing and stemming (using the Krovetz stemmer) words. As is common in sentiment studies (Das and Chen, 2007; Na et al., 2004; Kennedy and Inkpen, 2006) we added new features representing negated words. For example, the phrase “not good” is represented as a token “not-good” rather than as two separate words. This resulted in 43,910 features.

We used L_1-normalization, dividing term frequency matrix by the number of total word appearances in each document, and followed with a square root transformation, turning the Euclidean distance to the Hellinger distance. This multinomial geometry outperforms the Euclidean geometry in a variety of text processing tasks, as described in (Lafferty and Lebanon, 2005; Lebanon, 2005b).

Building a model solely based on the engineered term frequency features ignores the structure of a sentence or paragraphs. Using richer sets of feature may improve our model further; however, our contribution is presenting the manifold of emotions. We will use richer feature, especially handling sentence structures, in later research.

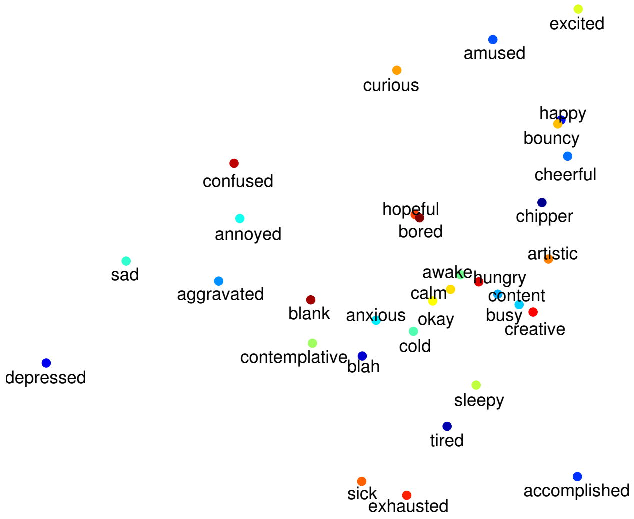

Figure 2: Mood centroids $\mathsf{E}(Z|Y = y)$ on the two most prominent dimensions in emotion space fitted from blog posts. The horizontal dimension corresponds to sentiments polarity and the vertical dimension corresponds to mental engagement level (compare with Figure. 1).

The document length histogram is close to an exponential distribution, with mean 113.51 words and standard deviation 146.65 words. There are plenty of short documents (520,436) having less than 50 words, but there are also some long documents (39,570) having more than 500 words. The average word length is 8.33 characters.

Two other datasets that we use in our experiments are the movie review data (Pang and Lee, 2005) and the restaurant review data[3] (Ganu et al., 2009) (using the same preprocessing described above).

4.2 Comparison with Psychological Models

In this section, we compare our model to Watson and Tellegen’s well known psychological model (Figure 1). Figure 2 shows the locations of mood centroids $\mathsf{E}(Z|Y = y)$ on the first two dimensions of the mood manifold. We make the following observations.

1. The horizontal axis expresses a sentiment polarity-like emotion. The left part features emotions such as `sad` and `depressed`, while the right part features emotions such as `accomplished`, `happy` and `excited`. This is in agreement with

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

[2] `http://www.lemurproject.org/`

[3] `http://www.cs.cmu.edu/~mehrbod/RR/`

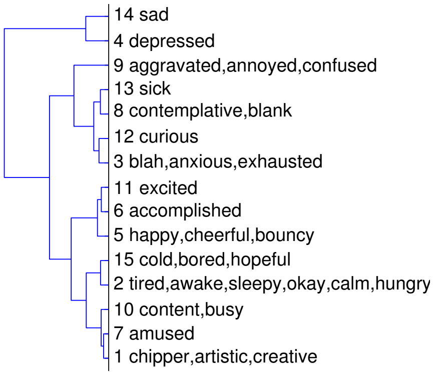

Figure 3: Dendrogram of moods using complete linkage function on Bhattacharyya distances between moods. The leaves are cut in 15 clusters to reduce clutters.

Watson and Tellegen's observations (see Figure 1) that identify sentiment polarity as the most prominent factor among human emotions.

2. The vertical axis expresses the level of mental engagement or energy level. The top part features emotions such as `curious` or `excited`, while the bottom part features emotions such as `exhausted` or `tired`. This agrees partially with the engagement dimension in the psychological model. However, the precise definition of engagement seems to be different. For example, in our model (Figure 2), high engagement imply active conscious mental states, such as `curious`, rather than passive emotions such as `astonished` and `surprised` (Figure 1).

3. The neutral moods `blank`, stay in the middle of the picture.

The mood centroid figure is largely intuitive, but the positions of a few centroids is somewhat unintuitive; for example `annoyed` has similar vertical location (energy level) as `bored`. We note, however, the manifold is higher dimensional and the dimensions beyond the first two provide additional positioning information.

It is interesting to consider the list of words that are most highly scored for each axis in our mood manifold. The words with highest weight associated with the horizontal axis (sentiment polarity) are: **`depress, sad, hate, cry, fuck, sigh, died`** on the left (negative) side and **`excite, awesome, yay, happy, lol, xd, fun`** on the right (positive) side. On the vertical axis (energy): **`tire, download, exhauste, sleep, sick, finishe, bed`** on the bottom side (low energy) and **`excite, amuse, laugh, not-wait, hilarious, curious, funny`** on the top side (high energy).

We conclude that there is in large part an agreement between the first two dimensions in our model and the standard psychological model. This agreement between our mood manifold and the psychological findings is remarkable in light of the fact that the two models used completely different experimental methodology (blog data vs. surveys).

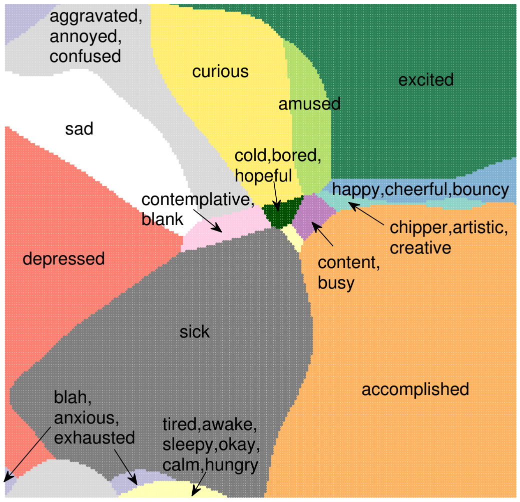

Figure 4: Tessellation of the space spanned by the first two dimensions of mood manifold with 15 "super-emotion" clusters ($\arg\max_y p(Z|Y=t)$).

4.3 Exploring the Emotion Space

Since emotion labels correspond to distributions $P(Z|Y)$, we can cluster these distribution in order to analyze the relationship between the different emotion labels. In the first analysis, we perform hierarchical clustering on the emotions in order to create emotional concepts of varying granularity. This is especially helpful when the original emotions are too fine, (consider for example the two distinct but very similar emotions `annoyed` and `aggravated`). In the second analysis we visualize the 2D tessellation corresponding to most likely emotions in mood space. This reveals additional information, beyond the centroid locations in Figure 2.

We use the Bhattacharyya dissimilarity,

$$D_B(f,g) = -\log \int \sqrt{f(z)g(z)}dz.$$

to measure dissimilarity between emotions, which corresponds to the log Hellinger distance between the underlying distributions. In the case of two multivariate Gaussians, it has the following closed form:

$$\begin{aligned} &D_B(N(\mu_1,\Sigma_1),N(\mu_2,\Sigma_2)) \\ &\quad = \frac{1}{8}(\mu_1-\mu_2)^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\mu_1-\mu_2) \\ &\qquad + \frac{1}{2}\log\left(\frac{\det((\Sigma_1+\Sigma_2)/2)}{\sqrt{\det\Sigma_1\det\Sigma_2}}\right). \end{aligned}$$

Following common practice, we add a small value to the diagonal of the covariance matrices to ensure invertibility.

Table 1: Macro F1 score and accuracy over the test set in multiclass emotion classification over top 32 moods(left) and over 7 clusters from Figure 3(right). Bold text represent statistically significant (t-test) improvement by using the mood manifold over the corresponding classification method in the original feature space.

		Original Space		Mood Manifold	
		F1	Acc.	F1	Acc.
LDA	full	n/a	n/a	**0.1247**	**0.1635**
	diag.	**0.1229**	0.1441	0.1160	**0.1600**
	spher.	0.0838	0.1075	**0.0896**	**0.1303**
QDA	full	n/a	n/a	**0.1206**	**0.1478**
	diag.	0.0878	0.0931	**0.1118**	**0.1463**
	spher.	0.0777	0.0989	**0.0873**	**0.1253**
Log.Reg.		0.1231	0.1360	**0.1477**	**0.1667**

		Original Space		Mood Manifold	
		F1	Acc.	F1	Acc.
LDA	full	n/a	n/a	**0.2800**	**0.3591**
	diag.	0.2661	0.2890	**0.2806**	**0.3504**
	spher.	0.2056	0.2252	**0.2344**	**0.2876**
QDA	full	n/a	n/a	**0.2506**	**0.3025**
	diag.	0.1869	0.1918	**0.2496**	**0.3088**
	spher.	0.1892	0.2009	**0.2332**	**0.2870**
Log.Reg.		**0.2835**	0.3459	0.2806	**0.3620**

Figure 3 shows the mood dendrogram obtained by hierarchical clustering of the top 32 emotions using the Bhattacharyya dissimilarity (complete linkage clustering). The bottom part of dendrogram was omitted due to lack of space. The clustering agrees with our intuition in many cases. For example,

1. `aggravated,annoyed` and `confused` are in the same tight cluster.
2. `sad` and `depressed` are very close cluster.
3. `happy`, `cheerful`, and `bouncy` are in the same tight cluster, which is close to `accomplished` and `excited`.
4. `tired, awake, sleepy, okay, calm` and `hungry` are in the same tight cluster.

The hierarchical clustering is useful in aggregating similar emotions. If the situation requires paying attention to one or two "types" of emotions, we can use a particular mood cluster to reflect the desired feature. For example, when analyzing product reviews we may want to partition the emotions into two clusters: positive and negative. When analyzing the effect of a new advertisement campaign we may be interested in a clustering based on positive engagement: excited / energetic vs. bored. Other situations may call for other clusters of emotions.

Figure 4 shows the tessellation corresponding to

$$f(z) = \arg\max_{y=1,\ldots,C} p(Z|Y=y).$$

For space and clarity purposes, we use 15 emotion clusters instead of the entire set of 32 emotions. The tessellation shows the regions being classified to each emotion cluster based only on the 2D space. We observe that:

1. As in Figure 2 the horizontal axis corresponds to negative(left) - positive(right) emotion and the vertical axis corresponds to energy level(or engagement): (top) `excited` and `curious` vs. (bottom) `tired` and `exhausted`.
2. The `depressed` region is spread significantly on the left-bottom side, and is neighboring the `sick` region and the `sad` region.
3. The region corresponding to the `happy, cheerful, bouncy` emotions neighbors the `accomplished` region and the `excited` region.

A similar tessellation of a higher dimensional Z space provides additional information. However, visualizing such higher dimensional spaces is substantially harder in paper format.

4.4 Classifying Emotions

One of the primary experiment in this paper is emotion classification. In other words, given a document x predict the emotion that is expressed in the text. As mentioned in the introduction, this classification can be done by constructing separate $p(y_i|x)$ models for every emotion (one-vs-all approach). However, the one vs. all approach is not entirely satisfactory as it ignores the relationships between similar and contradictory moods. For example, documents labeled as `sleepy` can be helpful when we fit a model for predicting `tired`. The mood manifold provides a natural way to incorporate this information, as documents from similar moods will be mapped to similar points on the manifold.

Besides testing different variants of LDA and QDA, we also compare logistic regression on the original input space and on the mood manifold (see Section 3.3).

4.4.1 Experiment Details

We performed emotion classification experiment (Table 1, left) on the Livejournal data. We considered the goal of predicting the most popular 32 moods. The class proportion varies in the range 1.72% to 6.52%.

Since 32 moods are too finer in practical usage, we designed coarser classification experiment (Table 1, right) using 7 clusters obtained by hierarchical clustering as in Figure 3. The task is to predict the 7 clusters and cluster proportion varies in the range 4.02% to 28.63%.

We also considered two binary classification tasks (Table 2) obtained by partitioning the set of moods into

Table 2: F1 and accuracy over test-set in sentiment polarity task (left): {cheerful, happy, amused} vs {sad, annoyed, depressed, confused}, and detecting energy level (right) {sick, exhausted, tired} vs. {curious, amused}. Bold text represent statistically significant (t-test) improvement by using the mood manifold over the corresponding classification method in the original feature space.

		Original Space		Mood Manifold	
		F1	Acc.	F1	Acc.
LDA	full	n/a	n/a	**0.7340**	**0.7812**
	diag.	0.7183	0.7436	**0.7365**	**0.7663**
	spher.	0.6358	0.6553	**0.7482**	**0.7699**
QDA	full	n/a	n/a	**0.6500**	**0.7446**
	diag.	0.6390	0.6398	**0.6704**	**0.7510**
	spher.	0.6091	0.6143	**0.7472**	**0.7734**
Log.Reg.		0.7350	0.7624	**0.7509**	**0.7857**

		Original Space		Mood Manifold	
		F1	Acc.	F1	Acc.
LDA	full	n/a	n/a	**0.7084**	**0.7086**
	diag.	0.6441	0.6449	**0.6987**	**0.6989**
	spher.	0.6343	0.6343	**0.6913**	**0.6913**
QDA	full	n/a	n/a	**0.5706**	**0.6100**
	diag.	0.6124	0.6413	**0.6268**	**0.6446**
	spher.	0.6239	0.6294	**0.6754**	**0.6767**
Log.Reg.		0.6694	0.6699	**0.7087**	**0.7089**

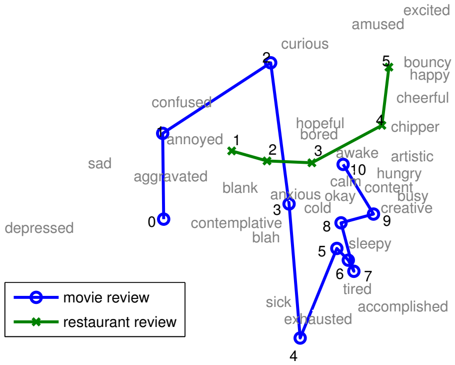

Figure 5: Projected centroids of each review score (higher is better) of movie reviews and restaurant reviews on the mood manifold. Both review start from the left side (negative sentiment in mood manifold) and continues to the right side (positive sentiment) with two different unique patterns.

two clusters (positive vs. negative clusters and high vs. low energy clusters). The class distributions of these binary tasks are 65.03% vs. 34.97% (sentiment polarity), and 52.17% vs. 47.83% (energy level)

We used half of the data for training and half for testing. To determine statistical significance, we performed t-tests on several random trials. Note that emotion prediction is a hard task, as similar emotions are hard to discriminate (consider for example discriminating between `aggravated` and `annoyed`). It is thus not surprising that prediction performances are relatively low, especially when discriminating between a large number of moods or clusters.

The LDA, QDA and L_2-regularized logistic regression models are implemented in MATLAB (the latter with LBFGS solver). We also regularized the LDA and QDA models by considering multiple models for the covariance matrices. We determined the regularization parameters by examining the performance of the model (on a validation set) on a grid of possible parameter values. We used the same parameters in all our experiments.

4.4.2 Classification Results

Table 1 and 2 compare classification results using the original bag of words feature space and the manifold model, using different types of classification methods: LDA, QDA with different covariance matrix models, and logistic regression. Bold faces are improvements over the baseline with statistical significance of t-test of random trials.

Most of experimental results show that the mood manifold model results in statistically significant improvements than using original bag of words feature. Improvements are consistent with various choices of classification methods: LDA, QDA, or logistic regression. The phenomenon is also persistent in variety of tasks: 32 mood classification, more practical 7 cluster classification, or binary tasks. Thus, introducing the mood manifold is indeed made the difference.

5 Application

5.1 Improving Sentiment Prediction using Mood Manifold

The concept of positive-negative sentiment fits naturally within our framework as it is the first factor in the continuous Z space. Nevertheless, it is unlikely that all sentiment analysis concepts will align perfectly with this dimension. For example, movie reviews and restaurant reviews do not represent identical concepts. In this subsection we visually explore these concepts on the manifold and show that the mood manifold leads to improved sentiment polarity prediction on these domains.

5.1.1 Sentiment Notion on the Manifold

We model a sentiment polarity concept as a smooth one dimensional curve within the continuous Z space. As we traverse the curve, we encounter documents corresponding to negative sentiments, changing smoothly into emotions corresponding to positive sentiments. We complement the stochastic embedding $p(Z|X)$

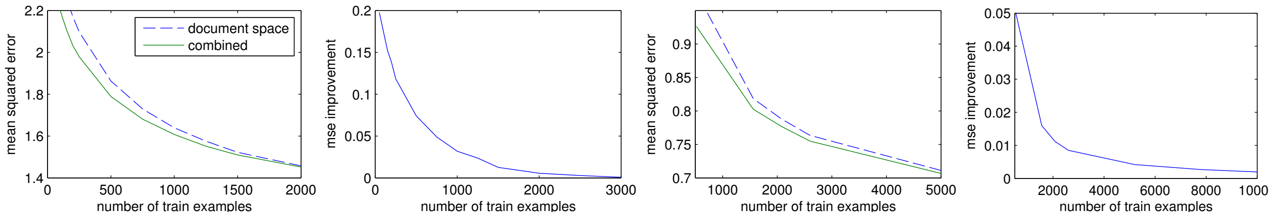

Figure 6: Test set mean squared error and its improvements on movie review (two figures on the left) and restaurant review (two figures on the right) as a function of the sentiment train set size. Prediction using the combined features outperforms the baseline (regression on document space) and the advantage is larger on smaller training set.

with a smooth probabilistic mapping $\pi(R|Z)$ into the sentiment scale. The prediction rule becomes

$$\hat{r} = \arg\max_{r} \int p(Z = z|X)\pi(R = r|Z = z)\, dz$$

and its approximated version is

$$\hat{r} = \arg\max_{r} \pi \left(R = r \middle| Z = \arg\max_{z} P(Z = z|X) \right)$$

Figure 5 shows the smooth curves corresponding to $\mathsf{E}\left[\pi(R = r|Z)\right]$ for movie reviews and restaurant reviews. Both curves progress from the left (negative sentiment) to the right (positive sentiment). But the two curves show a clear distinction: the movie review sentiment concept is in the bottom part of the figure, while the restaurant review sentiment concept is in the top part of the figure. We conclude that positive restaurant reviews exhibit a higher degree of excitement and happiness than positive movie reviews.

5.1.2 Improving Sentiment Prediction

The mood manifold captures most of the information for predicting movie review scores or restaurant review scores. Some useful information for review prediction, however, is not captured within the mood manifold. This applies in particular to phrases that are relevant to the review scores, and yet convey no emotional contents. Examples include (in the case of movie reviews) `Oscar`, `Shakespearean`, and `$300M`.

We thus propose to combine the bag of words TF representation with the mood manifold within a linear regression setting. We regularize the model using a group lasso regularization (Yuan and Lin, 2006), which performs implicit parameter selection by encouraging sparsity

$$\arg\min_{w} \frac{1}{n} \sum_{i=1}^{n} (w_1^T x^{(i)} + w_2^T z^{(i)} - y^{(i)})^2$$
$$+\lambda(||w_1||_2 + \lambda_2||w_2||_2).$$

Above, $z^{(i)}$ is the projection of $x^{(i)}$ on the mood manifold, and λ and λ_2 are regularization parameters. The regularization parameters was determined on performance on validation set and fixed throughout all experiments.

Figure 6 shows the test L_2 prediction error of our method and baseline (ridge regression trained on the original TF features) as a function of the train set size. The group lasso regression performs consistently better than regression on the original features. The advantage obtained from the mood manifold representation decays with the train set size, which is consistent with statistical theory. In other words, when the train set is relatively small, the mood manifold improves sentiment prediction substantially.

We also compared sentiment prediction using the bag of words features and sentiment prediction using the mood manifold exclusively. The mood manifold regression performs better than bag of words regression for smaller train set sizes but worse for larger train set sizes.

6 Summary and Discussion

In this paper, we introduced a continuous representation for human emotions Z and constructed a statistical model connecting it to documents X and to a discrete set of emotions Y. Our fitted model bears close similarities to models developed in the psychological literature, based on human survey data. The approach of this paper may also be generalized sequentially e.g., (Mao and Lebanon, 2007, 2009) and geometrically e.g., (Lebanon, 2003, 2009; Lebanon and Lafferty, 2004; Lebanon, 2005a,c; Dillon et al., 2007).

Several attempts were recently made at inferring insights from social media or news data through sentiment prediction. Examples include tracking public opinion (O'Connor et al., 2010), estimating political sentiment (Taddy, 2010), and correlating sentiment with the stock market (Gilbert and Karahalios, 2010). It is likely that the current multivariate view of emotions will help make progress on these important and challenging tasks.

Acknowledgments

The authors thank Joonseok Lee, Jeffrey Cohn, and Bo Pang for their comments on the paper. The research was funded in part by NSF grant IIS-0906550.

References

E.C.O. Alm. *Affect in text and speech*. ProQuest, 2008.

S. Das and M. Chen. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, 2007.

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

G. Ganu, N. Elhadad, and A. Marian. Beyond the stars: Improving rating predictions using review text content. In *12th International Workshop on the Web and Databases*. Citeseer, 2009.

M. Généreux and R. Evans. Distinguishing affective states in weblog posts. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, pages 40–42, 2006.

E. Gilbert and K. Karahalios. Widespread worry and the stock market. In *Proceedings of the International Conference on Weblogs and Social Media*, pages 229–247, 2010.

L. Holzman and W. Pottenger. Classification of emotions in internet chat: An application of machine learning using speech phonemes. Technical report, Technical Report LU-CSE-03-002, Lehigh University, 2003.

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

F. Keshtkar and D. Inkpen. Using sentiment orientation features for mood classification in blogs. In *IEEE International Conference on Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009*, pages 1–6, 2009.

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

R.J. Larsen and E. Diener. Promises and problems with the circumplex model of emotion. *Review of Personality and Social Psychology*, 13(13):25–59, 1992.

G. Lebanon. Learning Riemannian metrics. In *Proc. of the 19th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2003.

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

G. Lebanon. *Riemannian Geometry and Statistical Machine Learning*. PhD thesis, Carnegie Mellon University, Technical Report CMU-LTI-05-189, 2005b.

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

G. Lebanon. Axiomatic geomtries for text documents. In P. Giblisco, E. Riccomagno, M. P. Rogantin, and H. P. Wynn, editors, *Algebraic and Geometric Methods in Statistics*. Cambridge University Press, 2009.

G. Lebanon and J. Lafferty. Hyperplane margin classifiers on the multinomial manifold. In *Proc. of the 21st International Conference on Machine Learning*. Morgan Kaufmann Publishers, 2004.

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

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

G. Mishne. Experiments with mood classification in blog posts. In *1st Workshop on Stylistic Analysis Of Text For Information Access*, 2005.

S.M. Mohammad and P.D. Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 59(000):1–24, 2011.

J. Na, H. Sui, C. Khoo, S. Chan, and Y. Zhou. Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. *Advances in Knowledge Organization*, 9:49–54, 2004.

T. Nakagawa, K. Inui, and S. Kurohashi. Dependency tree-based sentiment classification using crfs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 786–794. Association for Computational Linguistics, 2010.

B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 122–129, 2010.

B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005.

B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2:1–135, 2008. ISSN 1554-0669.

C. Quan and F. Ren. Construction of a blog emotion corpus for chinese emotional expression analysis. In *Proceedings of the 2009 Conference on Empirical*

Methods in Natural Language Processing: Volume 3-Volume 3, pages 1446–1454. Association for Computational Linguistics, 2009.

V. Rubin, J. Stanton, and E. Liddy. Discerning emotions in texts. In *The AAAI Symposium on Exploring Attitude and Affect in Text (AAAI-EAAT)*, 2004.

J.A. Russell. Affective space is bipolar. *Journal of personality and social psychology*, 37(3):345, 1979.

J.A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

P. Shaver, J. Schwartz, D. Kirson, and C. O'connor. Emotion knowledge: Further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061, 1987.

R. Socher, J. Pennington, E.H. Huang, A.Y. Ng, and C.D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 151–161. Association for Computational Linguistics, 2011.

C. Strapparava and R. Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560. ACM, 2008.

C. Strapparava and A. Valitutti. Wordnet-affect: an affective extension of wordnet. In *Proceedings of LREC*, volume 4, pages 1083–1086. Citeseer, 2004.

M. Taddy. Inverse Regression for Analysis of Sentiment in Text. *Arxiv preprint arXiv:1012.2098*, 2010.

A. Tellegen, D. Watson, and L.A. Clark. On The Dimensional and Hierarchical Structure of Affect. *Psychological Science*, 10(4):297–303, 1999. ISSN 1467-9280.

D. Watson and A. Tellegen. Toward a consensual structure of mood. *Psychological bulletin*, 98(2):219–235, September 1985. ISSN 0033-2909.

D. Watson, L.A. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

M. Yik, J. Russell, and J. Steiger. A 12-point circumplex structure of core affect. *Emotion*, 11(4):705, 2011.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 2006.