This Text Has the Scent of Starbucks: A Laplacian Structured Sparsity Model for Computational Branding Analytics

William Yang Wang
School of Computer Science
Carnegie Mellon University
ww@cmu.edu

Edward Lin and John Kominek
Voci Technologies, Inc.
Pittsburgh, PA 15217
{ed.lin,john.kominek}@vocitec.com

Abstract

We propose a Laplacian structured sparsity model to study computational branding analytics. To do this, we collected customer reviews from Starbucks, Dunkin' Donuts, and other coffee shops across 38 major cities in the Midwest and Northeastern regions of USA. We study the brand related language use through these reviews, with focuses on the brand satisfaction and gender factors. In particular, we perform three tasks: automatic brand identification from raw text, joint brand-satisfaction prediction, and joint brand-gender-satisfaction prediction. This work extends previous studies in text classification by incorporating the dependency and interaction among local features in the form of structured sparsity in a log-linear model. Our quantitative evaluation shows that our approach which combines the advantages of *graphical modeling* and *sparsity modeling* techniques significantly outperforms various standard and state-of-the-art text classification algorithms. In addition, qualitative analysis of our model reveals important features of the language uses associated with the specific brands.

1 Introduction

In marketing science, branding is a modern marketing strategy of creating a unique image for a product in the customers' mind. Establishing the brand in the broad social context is just as important as building a good product (Makens, 1965; Lederer and Hill, 2001; Kim et al., 2013). In fact, blind taste test experiments have frequently shown how branding directly leads to the success of products and companies. Most notably is a continued study sponsored by Pepsi, known as the Pepsi Challenge[1], where Pepsi demonstrates how even though people preferred the taste of Pepsi, Coca-Cola's branding has made it more popular. Even now, Microsoft uses similar blind taste tests[2] to compare search engines, Bing and Google, showing that although participants prefer Bing's results, Google's brand might have strengthened over the years. These studies all suggest that brand and its associations play important roles in the customers' perceptions and decisions.

To accommodate the market change, companies frequently adjust branding strategies by analyzing how their customers receive and respond to branding messages. So far, such analysis is often done by using surveys and focus groups (Moon and Quelch, 2006), which is expensive and not time-efficient. Recently, with the advance of machine learning techniques, researchers from the chemistry and vision communities started to pay attention to the problem of automatic brand identification from smell (Luo et al., 2004) and images (Pelisson et al., 2003). In contrast, even though textual data that contains hidden branding information is abundantly available in many forms over the Web, automatic discovery and computational analysis on such data are not well studied in the past.

Computational branding analytics (CBA) seeks to extract information, trends, and demographics about a brand on the basis of free-form text, e.g. from blogs, Twitter comments, reviews, or forum posts. As described in Section 3, in this study we use a sub-

[1] http://en.wikipedia.org/wiki/Pepsi_Challenge
[2] http://www.bingiton.com/

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1325–1336,
Seattle, Washington, USA, 18-21 October 2013.

set of online Yelp reviews that discuss coffee shops. The main reason is that this source has the advantage of providing ground truth of multi-labeled data: each review has meta-information defining a 5-star rating, the object of the review, and the reviewer's name (from which we infer gender). For the purpose of this paper we decompose CBA into three sub-problems.

- How well can the brand being discussed be identified by the raw text?
- How well can the joint value of brand and rating be predicted?
- How well can the joint value of brand, rating, and gender be predicted?

There are two reasons why one may want to construct text-based classifiers of brand, rating, and gender, when such information is present in the review header. The first is that trained classifiers can then be applied to other data sources, such as blogs, where what is available is only the review itself. The second is that by "opening the hood" to the classifier one can examine which words exhibit high affiliation with the predicted variables. This can be done, for example, to contrast the preferences of males and females with respect to evaluating the qualities of a coffee shop. Examples of such insights are provided in Section 5.5.

In this paper, we propose a Laplacian structured sparsity model for computational branding analytics. Our main contributions are two-fold: first, in the novel task of automatic brand identification from text, we show that by incorporating the dependency structure and graphical interactions among local features, our model significantly outperforms various text classification algorithms such as the standard logistic regression, principle component analysis (PCA), linear kernel support vector machine (SVM), sparse, non-sparse, and mixed-penalty log-linear models. These improvements could also be seen from a joint brand-satisfaction prediction task and a gender-specific joint brand-satisfaction prediction task. In addition, our Laplacian augmented L_1-ball projection experiment shows that the advantage of Laplacian structured sparsity is robust across different parameter settings in a L_1-constrained problem. Secondly, the qualitative analysis of our machine learning model shows the interesting features and language use that relate to brand and its associated pragmatics.

In the next section, we outline related work in CBA, sparsity, and spectral graph learning. In Section 3, we describe the corpus in this study. The Laplacian structured sparsity model is introduced in Section 4. The experimental setup and results are presented in Section 5. A short discussion is followed in Section 6 and we conclude in Section 7.

2 Related Work

Early work on statistical brand analysis in the marketing community dates back to the work of Kuehn (1962), where he first hypothesizes that brand choice could be described as a learning process. Guadagni and Little (1983) further empirically tested the hypothesis by building a calibrated multinomial logistic regression model to predict the purchase of ground coffee, using the data from the optical scanning of product code in supermarkets. Outside the marketing community, statistical brand analysis is rarely seen. More recently, a study (Luo et al., 2004) applies neural networks to identify cigarette brands, with the hope of detecting illegal cigarettes from smell features. In image processing, researchers have studied the problem of brand identification from image using histogram comparison (Pelisson et al., 2003). However, to the best of our knowledge, even though textual data is vastly available, the problems of automatic brand identification from raw text and computational branding analytics, are new.

Although the domain of our data is on branding, our work also aligns with previous work in text and language classification. Over the years, logistic regression and linear kernel SVM have shown to be very successful in various regression and classification tasks in NLP (Chahuneau et al., 2012; Biadsy et al., 2011). Recently, sparse discriminative methods that model the sparse nature of text become attractive, because unlike dense models, they are less likely to overfit to the training data, easier to interpret, and often lead to state-of-the-art results. For example, Eisenstein et al. (2011b) use the $L_{1,\infty}$ sparsity model to discover sociolinguistic patterns. Wang et al. (2012a) compare lasso, ridge, and elastic net models to predict impoliteness behaviors in teenager conversations. Martins et al. (2011) investigate the tree-structured overlapping group lasso for

structured prediction problems. Chen et al. (2013) study the use of element-wise, group-wise, and hierarchical sparsity models for dialogue act classification. Sparse inducing priors are also investigated and shown to be effective in generative models for topic modeling (Eisenstein et al., 2011a; Wang et al., 2012b; Paul and Dredze, 2012).

Besides lacking sparsity, since the traditional discriminative methods in NLP often use interdependent features such as n-grams tokens, and part-of-speech tags, they also suffer from the problem of not explicitly modeling the complex dependency structure and interaction of local features from a global perspective. To solve this problem, graph methods seem to be a good solution, because they are simple, generalizable, and are often used to model such complex dependency structures (Cohen, 2012). However, combining the sparse modeling and spectral graphical modeling approaches in a principled way is challenging. Belkin et al. (2006) and Weinberger et al. (2007) are among the first to investigate graph Laplacians as a manifold regularization method for statistical learning. Recently, Gao et al. (2012) propose a histogram intersection based kNN method to construct a Laplacian matrix for a least-square sparse coding problem in image processing. Unfortunately, this method might be too specific to the SIFT-based image coding tasks, thus might not be applicable to the text classification problem that utilizes n-gram lexical features.

3 Datasets

We collected Yelp reviews from 1,860 Starbucks, Dunkin' Donuts[3], and other coffee shops all over the Midwest and Northeast regions in the period of 2009. A detail statistics of our data can be found in Table 1. The Midwest region includes 12 states[4] and 19 major cities, and the Northeast region includes 9 states[5] and 19 major cities. For each region, we divide the coffee shops into 60% training, 20% development, and 20% test, and there are no overlaps of coffee shops among these subsets. There are three values for the brand label: Starbucks, Dunkin' Donuts, and all other coffee shop brands. The majority class is "all other coffee shop brands", and the majority baseline is shown in Table 2. In the task of joint brand-satisfaction prediction, we utilize the review scores to approximate user satisfaction: scores 1-2 as the unsatisfactory label, 3 as moderate, and 4-5 as satisfactory. Since the Yelp reviews do not reveal the reviewer's gender, we use a similar method that U.S. Census Bureau used (OConnell and Gooding, 2006): we first automatically match the first name of the reviewer with the prior name-gender distributions in the census records, then manually examine the no-match cases and a subsample of the matched cases. For those who we cannot determine the gender, the review will be dropped from the gender-specific brand-satisfaction prediction task. After filtering, there are 8,528 documents for training, 2,928 for development, and, 3,046 for testing. Since the focus of this paper is not on feature engineering, we use unigram features to represent each review. Below is an example of positive review from a male Starbucks customer from Midwest.

	Coffee Shops			Reviews		
	Train	Dev.	Test	Train	Dev.	Test
1	451	150	150	3,513	1,087	1,424
2	665	222	222	6,982	2,530	2,358
T.	1,116	372	372	10,495	3,617	3,782

Table 1: Dataset statistics. *1: midwest region. 2: northeast region. T.: total.*

> *My favorite place for my iced vanilla lattes. They have screwed up my order before: instead of a grande, I got a venti. Not a fan of their pastries though. I got a donut once, and ended up feeding it to a pigeon in city garden. Friendly and fast service. Not open Sundays.*

The coffee shop dataset is freely available[6] for research purposes.

4 Our Approach

4.1 Problem Formulation and Predictive Tasks

The automatic brand identification problem could be considered as a traditional multiclass classifica-

[3]We chose these two brands because they are reported as the leading coffee shops by WSJ (Ovide, 2011) and Forbes (DiCarlo, 2004).

[4]IL, WI, SD, ND, MN, MO, OH, NE, KS, IA, IN, and MI.

[5]CT, ME, MA, NH, RI, VT, NJ, NY and PA.

[6]http://www.cs.cmu.edu/~yww/data/emnlp2013.zip

tion problem where the estimated label $\hat{Y}$ could be drawn from $Mult(\Gamma)$, where Γ is the parameter for the multinomial distribution. To solve this, a simple but accurate solution is to decompose the multiclass problem into multiple binary classification problems (Rifkin and Klautau, 2004) by training k one-vs-all binary classifiers, and then use the argmax criteria to select the best hypothesis from the k posteriors. As for a binary classifier, we need to infer the posterior from a Bernoulli distribution that is parametrized by $\hat{\theta}_y$. Similarly, we can derive k binary classifiers:

$$\hat{\theta}_y^{(1)}, \hat{\theta}_y^{(2)}, ..., \hat{\theta}_y^{(k)}. \quad (1)$$

So, instead of drawing $\hat{Y}$ from a multinomial distribution $Mult(\Gamma)$, we can draw the final label $\hat{Y}$ that has the largest posterior across all k classifiers:

$$\hat{Y} = \underset{Y,i=1,2,...,k}{\operatorname{argmax}} \Pr(Y|\hat{\theta}_y^{(i)}, \vec{X}_t) \quad (2)$$

where $\vec{X}_t$ is the testing vector, and $\Pr(Y|\hat{\theta}_y^{(i)}, \vec{X}_t)$ is the posterior probability given the learned classifiers and the testing vector.

In this paper, we investigate three multiclass classification tasks: first, we perform a 3-way classification task for automatic brand identification. In the task of brand-satisfaction prediction, we model the brand and the satisfaction label at the same time (Chahuneau et al., 2012): we perform the task of jointly predicting aggregate brand-satisfaction score for a review using 9-way classification. Similarly, we perform 18-way classification for the gender-specific joint brand-satisfaction prediction task.

4.2 The Log-Linear Framework and Its Regularized Variants

If we consider the standard logistic regression model as the binary classifier in this log-linear framework[7], then each classifier can be written as:

$$\hat{\theta}_y = \frac{\exp\left(\vec{W}^\top \vec{X}_j\right)}{1 + \exp\left(\vec{W}^\top \vec{X}_j\right)} \quad (3)$$

here, $\vec{X}_j$ is the j-th observed feature vector, label $y \in \{0, 1\}$, and $\vec{W}$ is a vector of the coefficients. To estimate the model parameters in equation (3), we only need to set the weights $\vec{W}$. We can obtain the following log likelihood, and its gradient function by taking the first-order partial derivative of $\vec{W}$:

$$\ell = \sum_j y_j \log \hat{\theta}_{y_j} + (1 - y_j) \log(1 - \hat{\theta}_{y_j}) \quad (4)$$

$$\frac{\partial \ell}{\partial \vec{W}} = \sum_j \left(\frac{\partial \hat{\theta}_{y_j}}{\partial \vec{W}}\right)\left(\frac{y_j}{\hat{\theta}_{y_j}} - \frac{1 - y_j}{1 - \hat{\theta}_{y_j}}\right) \quad (5)$$

$$\frac{\partial \hat{\theta}_{y_j}}{\partial \vec{W}} = \left(\hat{\theta}_{y_j} - (\hat{\theta}_{y_j})^2\right)\vec{X}, \quad (6)$$

since the log likelihood objective function (4) is concave, using standard gradient ascent with maximum likelihood estimation can solve the problem. However, this model does not penalize the noisy features and unreliable features that might overfit to the training data. To address this issue, we introduce the L_1 norm from lasso technique (Tibshirani, 1996) to regularize the above likelihood function. Thus, instead of maximizing the likelihood, we can minimize the loss function of the negative log-likelihood with a linear penalty:

$$\min\Big(-\ell + \lambda_1 ||\vec{W}||\Big) \quad (7)$$

where λ_1 is the regularization coefficient. The benefit of L_1 penalty in a discriminative model is similar to the double exponential distribution of the sparse priors in generative models (Eisenstein et al., 2011a): they both push the weights of many noisy features into zeros, revealing only the important features. However, since the L_1 penalty can introduce discontinuities to the original convex function, we can also consider an alternative non-sparse ridge estimator (Le Cessie and Van Houwelingen, 1992) with log loss and L_2 norm, and has the convex property:

$$\min\Big(-\ell + \lambda_2 ||\vec{W}||^2\Big) \quad (8)$$

Another option that balances the sparsity and smoothness would be the elastic net model (Zou and Hastie, 2005) that uses the composite penalty:

$$\min\Big(-\ell + \lambda_1 ||\vec{W}|| + \lambda_2 ||\vec{W}||^2\Big) \quad (9)$$

4.3 The Laplacian Structured Sparsity Model

So far, none of the above element-wise penalty models in the previous subsection takes into account the

[7] We thank Jacob Eisenstein for the initial derivation of the logistic regression model.

dependency structure of the local features. Inspired by Gao et al.(2012), we group the local features that have similar distributions together. The intuition is that, for features that have very similar empirical distributions in the training set, their weights should not be drastically different after the learning process in the same task. In our new objective function, it is desirable to introduce a new component that structurally penalize these cases where features that are very similar to each other, but have learned completely different weights, probably due to the noise or the data sparsity issue in the training data.

The Objective Function: To do this, we first define an inter-feature affinity matrix A, where $A_{(p,q)}$ measures the similarity between a pair of features p and q. In the spectral graph theory, this affinity matrix can be viewed as a weighted undirected graph $G = (V, E)$, where each node V_p denotes a feature p, and each edge $E_{(p,q)}$ indicates the closeness among the features p and q. We also introduce a weighted diagonal degree matrix D, of which each element in the diagonal $D_{(p,p)}$ is the sum of all weighted connections of node V_p: $D_{(p,p)} = \sum_{q=1}^{Q} A_{(p,q)}$. We propose the following objective function:

$$\min \Big(-\ell + \lambda_1 ||\vec{W}|| + \lambda_2 ||\vec{W}||^2 \tag{10}$$

$$+ \alpha \sum_{(p,q)} ||\vec{W}_p - \vec{W}_q||^2 A_{(p,q)} \Big) \tag{11}$$

We then denote a graph Laplacian matrix $L = D - A$ (Belkin and Niyogi, 2001), and rewrite the objective function as:

$$\min \Big(-\ell + \lambda_1 ||\vec{W}|| + \lambda_2 ||\vec{W}||^2 \tag{12}$$

$$+ \alpha (\vec{W}^\top L \vec{W}) \Big) \tag{13}$$

where α is the regularization parameter for the Laplacian structured sparsity term. Intuitively, the objective function can be interpreted as the sum of a negative log loss function, the sparsity-inducing penalty, the quadratic penalty, and the Laplacian structured penalty. Or, another view of this new model could be seen as a Laplacian augmented elastic net model where structured sparsity and feature interaction are considered.

The Laplacian Matrix: In this model, a key aspect is to derive the Laplacian matrix L. We propose the following three steps to learn the Laplacian matrix:

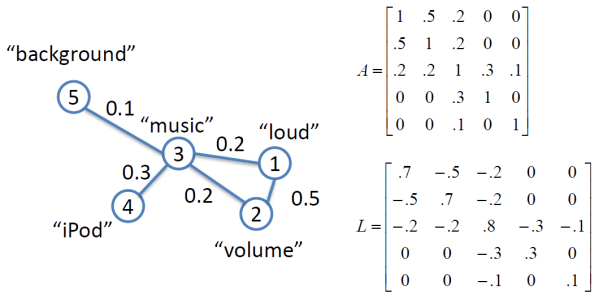

Figure 1: An example of the graph G, the corresponding affinity matrix A, and the corresponding Laplacian matrix L.

1. **Construct the distance matrix** $Dist$**.** To construct the distance matrix between each feature, we first transpose the instance-feature matrix, $I = \sum_j \vec{X}_j$, and assume that each feature (e.g. unigram in our task) is a random variable that has a multinomial distribution over the instances in the training set. Then, we compare each pair of features, and calculate the inter-feature distance matrix $Dist$ with Euclidean distance as a measure, and use the k-nearest neighbors (kNN) method (Beyer et al., 1999) to select the top neighbors of each feature.

2. **Derive the affinity matrix** A**.** To assign the weight on the edge $E_{(p,q)}$ for each connected nodes (the kNN of V in $Dist$), we use the cosine similarity $cosine(V_p, V_q)$ metric (Wang and Hirschberg, 2011).

3. **Generate the degree matrix** D **and Laplacian matrix** L**.** As discussed earlier, we sum up the symmetric affinity matrix by row, and obtain a diagonal degree matrix D, and we further define a Laplacian matrix $L = D - A$.

To calculate the above matrices in an efficient manner, we partition the covariate into blocks, and process each block in parallel (Chen et al., 2011). An intuitive example of the graph G, its associated affinity matrix A, and Laplacian matrix L, is shown in Figure 1.

Parameter Estimation: Regarding the optimization of objective function in (12-13), a notable problem is that the sparsity inducing L_1 term is non-differentiable, whereas this is not the case for the L_2 norm and the Laplacian structured sparsity term. If we first take the derivative of the latter two terms,

and we can derive the following gradient components:

$$\frac{\partial(\lambda_2||\vec{W}||^2 + \alpha(\vec{W}^\top L\vec{W}))}{\partial\vec{W}} \tag{14}$$

$$= 2\lambda_2\vec{W} + \alpha(\vec{W}^\top L^\top + \vec{W}^\top L) \tag{15}$$

$$= 2\lambda_2\vec{W} + \alpha(L^\top + L)\vec{W} \tag{16}$$

since our Laplacian matrix is symmetric, we can rewrite (16) as

$$2(\lambda_2\vec{W} + \alpha L\vec{W}) \tag{17}$$

Then, we combine the gradient of the log loss function in (5) with (17), and apply a bound-constrained re-formulation (Schmidt et al., 2007) and the limited memory BFGS (L-BFGS) method (Liu and Nocedal, 1989) to solve the L_1 regularized problem. The L-BFGS method has relatively low space complexity, and does not require the calculation of full Hessian matrix, thus it is often used for L_1 optimization problems.

Augmented Laplacian for an L_1-Constrained Problem: Instead of formulating the L_1-regularized problem by adding the L_1 norm, an alternative solution is to formulate a L_1-constrained problem by fixing the sum of all weights τ in the weight vector $\vec{W}$. The reason is because adding the L_1 norm will make the objective function not continuously differentiable, where as the L_1 constraint could be just a simple linear constraint (Lee et al., 2006). Thus, the alternative L_1-constrained problem could be defined as:

$$\min(-\ell), s.t. \sum_p \vec{W}_p \leq \tau \tag{18}$$

To test the robustness of Laplacian structured sparsity term in the setup of a L_1-constrained problem, we can incorporate the Laplacian penalty term into the above formula, and derive:

$$\min\Big(-\ell + \alpha(\vec{W}^\top L\vec{W})\Big), s.t. \sum_p \vec{W}_p \leq \tau \tag{19}$$

Note that the Laplacian matrix is positive-semidefinite,

$$\vec{W}^\top L\vec{W} = \vec{W}^\top \sum_{(p,q)} L_{(p,q)}\vec{W} \tag{20}$$

$$= \sum_{(p,q)} \vec{W}^\top L_{(p,q)}\vec{W} \tag{21}$$

$$= \sum_{(p,q)} ||\vec{W}_p - \vec{W}_q||^2 A_{(p,q)} \tag{22}$$

because this graph Laplacian penalty can be viewed as a quadratic term, and the objective function in equation 19 is now convex differentiable and will produce sparse estimates, so that we are able to use a limited-memory projected quasi-Newton method (Schmidt et al., 2009) to solve the dual form of this problem. The Lagrangian dual form of the problem in equation 19 can be written as:

$$\mathcal{L}(\vec{W}, \xi) = -\ell + \alpha(\vec{W}^\top L\vec{W}) \tag{23}$$

$$+ \beta\left(\sum_p \vec{W}_p - \tau\right) - \xi\vec{W} \tag{24}$$

where $\beta \in \mathbb{R}$ is a Lagrange multiplier, and $\xi \in \mathbb{R}^p_+$ is a p-dimensional vector of non-negative Lagrange multipliers. And then we can take first-order partial derivative with regard to $\vec{W}$, and set it to zero to derive the optimality:

$$\frac{\partial\mathcal{L}}{\partial\vec{W}} = -\sum_j \left(\hat{\theta}_{y_j} - (\hat{\theta}_{y_j})^2\right)\vec{X}\left(\frac{y_j}{\hat{\theta}_{y_j}} - \frac{1-y_j}{1-\hat{\theta}_{y_j}}\right) \tag{25}$$

$$+ 2\alpha L\vec{W} + \beta - \xi = 0 \tag{26}$$

To speed up the training, we use the linear-time L_1-ball projection method from Duchi et al. (2008) in our implementation.

5 Experiments

We first compare our model to various baselines in the 3-way automatic brand identification task. Besides the logistic regression, lasso, ridge and elastic net model that we introduced in Section 4.2, we also compare with a PCA-based logistic regression model where the dimensions of the feature space is reduced in half before the classification. A state-of-the-art linear kernel SVM model (Chang and Lin, 2011) is also taken into the comparison. In the second part, we perform 9-way joint classification of the brand-satisfaction labels. Similarly, we also perform a 18-way joint classification of the brand-gender-satisfaction labels. To test the robustness of our model, we vary the levels of sparsity of our Laplacian augmented method in a L_1-constrained problem. Finally, we analyze the identified features for CBA. Throughout this section, we use classification accuracy to report the results. We tune the regularization parameters of log-linear models and

Method	Dev.	Test
Majority class	75.67	78.08
Logistic regression	91.98	91.06
Linear SVM	92.45	91.75
PCA	91.67	91.20
Lasso	92.81	91.96
Ridge	92.56	91.67
Elastic net	92.81	91.83
Laplacian structured sparsity	**93.17***	**92.44***

Table 2: The automatic brand identification (3-way) performances. The best result is highlighted in **bold**. * indicates $p < .001$ comparing to the second best result.

Method	Dev.	Test
Majority class	55.43	55.18
Logistic regression	65.80	65.80
Linear SVM	67.67	65.44
PCA	63.92	62.53
Lasso	68.37	66.84
Ridge	67.79	65.55
Elastic net	68.79	66.82
Laplacian structured sparsity	**69.56***	**67.32***

Table 3: The joint brand-satisfaction prediction (9-way) performances. The best result is highlighted in **bold**. * indicates $p < .001$ comparing to the second best result.

the cost parameter of the SVM on the development set, and report results on both the development set and the held-out test set. The parameter for kNN was set to 5 according to previous literature (Gao et al., 2012). A paired two-tailed t-test is used to test the statistical differences among various models.

5.1 Automatic Brand Identification from Text

Given any piece of raw text from the Web (e.g. blogs, tweets, news, or forum posts), the first task for CBA is to identify which brand this text is related to. Our customer review data set is useful for this task, because the ground truth of the brand label is attached to each review. Table 2 shows the result of our model in this automatic brand identification task. In this 3-way classification task, the overall results indicate that it is relatively easy to identify the related brand from customer reviews. When evaluating our Laplacian structured sparsity model, our proposed model obtains the best performances of 93.17% and 92.44%, which are statistically better than the second best results ($p < .001$) in both datasets.

5.2 Joint Brand-Satisfaction Prediction

In our training data set, we observe a subtle correlation between the brand and satisfaction labels ($r = 0.09$, $p < .001$), which suggests us that it might be interesting to perform a joint prediction task for the brand-satisfaction labels. This task is also attractive from the business perspective, because it would be very useful for the companies to directly identify user's level of satisfaction about their brands. Table 3 shows that we achieve 69.56% accuracy on the development set, and 67.32% accuracy on the test set using our proposed Laplacian structured model ($p < .001$ comparing to the second best results).

Method	Dev.	Test
Majority class	28.24	27.68
Logistic regression	36.03	35.16
Linear SVM	41.05	39.49
PCA	35.35	34.44
Lasso	40.74	39.53
Ridge	40.98	38.94
Elastic net	41.15	38.96
Laplacian structured sparsity	**41.22***	**40.22***

Table 4: The joint brand-gender-satisfaction prediction (18-way) performances. The best result is highlighted in **bold**. * indicates $p < .001$ comparing to the second best result.

5.3 Joint Brand-Gender-Satisfaction Prediction

Another big interest in the marketing community is to predict subgroup preferences of specific brands. In this direction, we perform a 18-way joint brand-gender-satisfaction prediction using the gender labels that we described in Section 3. Table 4 shows that our proposed Laplacian structured sparsity model obtains a test accuracy of 40.22%, significantly better than the second best result ($p < .001$).

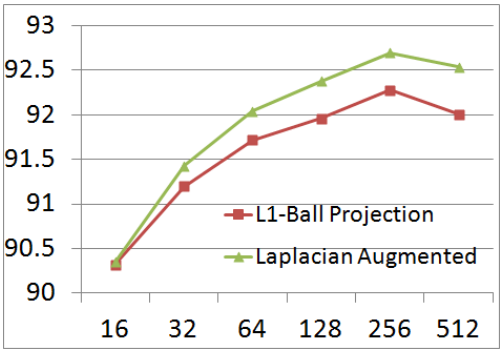

Figure 2: Automatic brand identification test performance varying the level of sparsity τ in a L_1 constrained problem.

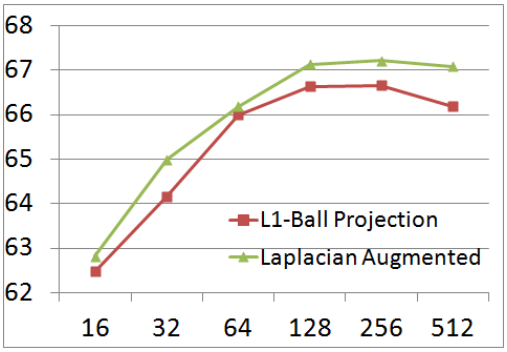

Figure 3: Joint brand-satisfaction prediction test performance varying the level of sparsity τ in a L_1 constrained problem.

5.4 Varying the Level of Sparsity in a L_1-Constrained Problem

To test the robustness of the Laplacian structured sparsity component, we exponentially increase the sum of weights τ to vary the level of sparsity in a L_1-constrained setup. When τ increases, the non-zero weights in the model also increases. Figures 2 and 3 show that the Laplacian augmented L_1-ball projection statistically outperform the L_1-ball projection baseline in all levels of sparsity ($p < .001$). In Figure 4, Laplacian augmented L_1-ball projection is also statistically better than the L_1-ball projection ($p < .001$), except when $\tau = 32$ and $\tau = 64$.

5.5 Exploratory Data Analysis

We outline the top 15 keywords from the Laplacian structured sparsity model that are associated with the Starbucks and Dunkin' Donuts brands in the automatic brand identification task in the Table 5. First of all, it is observed that our model has discovered synonyms for both brands: "sbux", "dd", "dds".

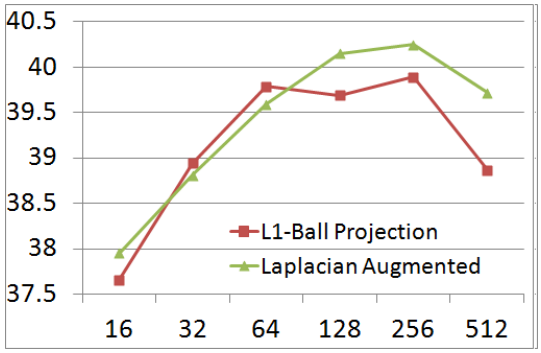

Figure 4: Joint brand-gender-satisfaction prediction test performance varying the level of sparsity τ in a L_1 constrained problem.

Also, the results imply that Starbucks' unique cup size branding strategy, "venti", "grande", "tall", has resonated with their customers as the words prominently show up as top features in reviews. Aligned with previous study in marketing science (Moon and Quelch, 2006), an informative set of features related to Starbucks store decorations showed up in our model: "store", "restroom", "public", "bathroom", and "spacious". In contrast, these features stopped to show up on the list of Dunkin' Donuts. Instead, TV and game (sports), which are indeed important features of dining at Dunkin' Donut, appeared. Note that Baskin-Robbins, which is a sub-brand of Dunkin' Brands Group, Inc., also appeared as informative features to predict Dunkin' Donuts.

To understand the preferences of different gender subgroups towards the two brands, we contrast in Table 6 and Table 7 the top features that identify the satisfied **female** and **male** customers in the joint brand-gender-satisfaction prediction task.

Interestingly, it seems that the female customers identify Starbucks as a place for "studying", with "fireplace" as the top preference of the spots in the store, and "winter" is also a high-ranked feature. Also, the adjective "super" was frequently mentioned by the female Starbucks customers (but not the males). As for Dunkin' Donuts, the top-ranked keywords are still mainly associated with its names, but it seems the snack "Munchkins" is highly preferred by the female customers. Not surprisingly, the cue words that the male customers identify the Starbucks brand do not always agree with those of the females. For example, instead of "fireplace", they prefer staying at the "patio", and drink the coffee from the "clover" brewing system. Interestingly,

Starbucks	weight	Dunkin'	weight
starbucks	1.9365	dd	2.4224
sbux	1.0152	dunkin	1.7781
venti	0.8216	donuts	1.6989
corporate	0.7032	dunks	1.6455
store	0.6580	dds	1.4936
particular	0.6512	donut	1.3979
tall	0.5496	dunkins	1.3729
restroom	0.5447	glazed	0.9975
tourists	0.5431	robbins	0.9402
public	0.5260	baskin	0.8578
lines	0.4956	sugar	0.6475
drink	0.4787	d	0.6327
bathroom	0.4721	ice	0.5835
spacious	0.4629	stale	0.5404
location	0.4611	game	0.5049
grande	0.4563	tv	0.5010

Table 5: Top features that identify the Starbucks and Dunkin' Donuts brands from the best model.

.

on the Dunkin's side, "munchkins" also disappeared and replaced by "glazed" (donuts). However, both males and females agreed that "fast" or "quick" service was an important feature of creating satisfaction, which echoes with the result from self-reported customer surveys (Moon and Quelch, 2006).

The word "**name**" is a prominent indicator for the female customers of Starbucks: at first we were puzzled, but after we digged into the database, we found reviews such as:

- *"... and the baristas are one of the nicest they always ask for your **name**, so you never end up with coffee meant for the guy behind you."*
- *"... she asked me my name and i told her and she excidetly proclaimed melissa and wrote my **name** on the cup. This place was probably one of the better starbucks ive been to."*
- *"... all of their employees are really friendly, and embarrassingly enough most know me by **name** and know my typical drink order grande nonfat misto with a flavor shot of white mocha. This is actually very helpful."*

The above examples show how our system effectively serves as a salient keyword spotter. And that as a keyword spotter one can use it to extract surrounding context and feed that through to the next stage of analysis, including examination by humans. This is extremely practical and useful, because it provides actionable items. For example, analysts can advise managers to revise their training manual and tell store employees to *remember the names of your frequent female customers*.

Starbucks	weight	Dunkin'	weight
starbucks	0.5013	dd	0.6931
chain	0.4268	dds	0.5620
winter	0.3382	dunkin	0.5344
fireplace	0.3089	donuts	0.4270
studying	0.2972	donut	0.3732
particular	0.2967	dunks	0.3687
super	0.2786	morning	0.3077
name	0.2543	quick	0.3012
know	0.2443	how	0.2940
because	0.2263	munchkins	0.2758

Table 6: Top features that jointly identify the satisfied **female** customers and the Starbucks and Dunkin' Donuts brands from the best model.

Starbucks	weight	Dunkin'	weight
starbucks	0.6632	dd	0.7491
throw	0.3514	dunkin	0.6075
know	0.2959	dds	0.5333
store	0.2885	donuts	0.5326
fix	0.2498	donut	0.3215
particular	0.2487	dunks	0.3158
sbux	0.2462	morning	0.3095
patio	0.2349	rush	0.3030
prefer	0.2324	fast	0.2979
clover	0.2215	moving	0.2520
corporate	0.2153	glazed	0.2326

Table 7: Top features that jointly identify the satisfied **male** customers and the Starbucks and Dunkin' Donuts brands from the best model.

6 Discussions

In our preliminary experiments, we have also experimented with the setup where the two keywords "starbucks" and "dunkin" were removed from the list of features. This change resulted in a uniformed 2% decrease in performances across all the models in Table 2, which did not affect the comparisons.

However, we kept these two keywords in our final experiments, because the reviewers sometimes mention "Dunkin" in Starbucks reviews, and vice versa. Removing the two keywords could be problematic, since it changes the natural distribution of the data.

Regarding the alternative problem setups, our preliminary experiments showed that instead of using one-vs-all binary classifiers, a direct 9-way multi-class classification of joint brand-satisfaction labels using logistic regression only resulted an accuracy of 62%. We also did not adopt the hierarchical classification pipeline, where instead of performing joint classification, multiple layers of classifiers could be trained to classify brand, gender, and satisfaction labels incrementally. This is because the hierarchical classifiers suffered from the error propagation problem, and the second/third layer classifier could not correct the errors from the previous layers (Bennett and Nguyen, 2009).

Our proposed method to generate inter-feature affinity matrix captures interesting dependency of features in this dataset. For example, although the words "frappuccino" and "slurping", "furniture" and "mismatched' are semantically very different, our method actually group them together due to the subtle interactions of these word pairs in our tasks. The example in Figure 1 is also very specific to our dataset. This is very useful, because the word semantic similarity might be context-dependent, and our method learns and adapts the semantic similarity on the fly, hinges on the particular training set. On the other end of the spectrum, even though our method is desirable in our task, one might need to be cautious when working on very small data sets with only a handful of samples. This is because small samples typically have large variances in feature distributions, and that the generated Laplacian matrix might not be as reliable as in our study. To alleviate this potential problem, one might consider building the Laplacian matrix using external resources such as WordNet or FrameNet, even though this approach could also introduce biases due to the mismatched task domains.

We also observed that the accuracy of the automatic brand identification task was high, indicating the promising future of CBA for hidden brand information from other genres of text over the Web. Although the performances of joint brand-satisfaction and joint brand-gender-satisfaction predictions are relatively lower, there is still much room for improvements: for example, using the syntactic, semantic, and meta-data features could potentially enrich the proposed model. Also, it is possible to consider the higher order n-gram features for better exploratory data analysis. However, since the focus of this paper is a proof of concept for Laplacian structured sparsity models and computational branding analytics, we have not yet explored various multi-view representations to augment our model.

Why does Laplacian structured sparsity model work better in these classication tasks? Similar to the application in image classifcation (Gao et al., 2010), one advantage of Laplacian regularization in text classification is that our model can explicitly model the dependency of local features. Another reason is the expressiveness of our model: our model allows one to express the feature interactions in a structured manner. Thirdly, by embedding the structure in the regularization term, our model is more flexible: one can now control the structured penalty by tuning the regularization parameter on the development set.

7 Conclusions

We introduce a Laplacian structured sparsity model for computational branding analytics (CBA). In the automatic brand identification, our model achieves the best result, dominating many competitive baselines. We also introduce the tasks of joint brand-satisfaction and brand-gender-satisfaction predictions, and show that the Laplacian structured sparsity do well in these tasks. A closer evaluation that varying the levels of sparsity in a L_1 constrained problem also indicates that the Laplacian augmented L_1-ball projection model can provide state-of-the-art results. By examining the weights of the derived Laplacian structured sparsity model, interesting indicators of brands and theirs gender-specific customer satisfaction associations are also discovered. In the future, we would like to investigate other methods for generating robust inter-feature Laplacians that include deep syntactic and semantic features.

Acknowledgement

The authors would like to thank the anonymous reviewers for valuable comments.

References

M. Belkin and P. Niyogi. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems (NIPS)*.

M. Belkin, P. Niyogi, and V. Sindhwani. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research (JMLR)*.

P. N. Bennett and N. Nguyen. 2009. Refined experts: improving classification in large taxonomies. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*.

K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. 1999. When is nearest neighbor meaningful? *Proceedings of the International Conference on Database Theory (ICDT)*.

F. Biadsy, W.Y. Wang, A. Rosenberg, and J. Hirschberg. 2011. Intoxication detection using phonetic, phonotactic and prosodic cues. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*.

V. Chahuneau, K. Gimpel, B.R. Routledge, L. Scherlis, and N.A. Smith. 2012. Word salad: Relating food prices and descriptions.

C.C. Chang and C.J. Lin. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*.

W. Y. Chen, Y. Song, H. Bai, C. J. Lin, and E. Y. Chang. 2011. Parallel spectral clustering in distributed systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Y. N. Chen, W. Y. Wang, and A. I. Rudnicky. 2013. An empirical investigation of sparse log-linear models for improved dialogue act classification. In *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*.

W. W. Cohen. 2012. Learning similarity measures based on random walks. In *Procceedings of the 21nd ACM International Conference on Information and Knowledge Management (CIKM)*.

L. DiCarlo. 2004. Dunkin' donuts vs. starbucks. In *Forbes.com - Monday Matchup*.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. 2008. Efficient projections onto the l1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning (ICML)*.

J. Eisenstein, A. Ahmed, and E. Xing. 2011a. Sparse additive generative models of text. *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*.

J. Eisenstein, N. A. Smith, and E. P. Xing. 2011b. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*.

S. Gao, I. W. Tsang, L. T. Chia, and P. Zhao. 2010. Local features are not lonely–laplacian sparse coding for image classification. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

S. Gao, I. Tsang, and L. Chia. 2012. Laplacian sparse coding, hypergraph laplacian sparse coding, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

P. M. Guadagni and J. D. C. Little. 1983. A logit model of brand choice calibrated on scanner data. *Marketing science*.

M. K. Kim, K. Lopetcharat, and M. A. Drake. 2013. Influence of packaging information on consumer liking of chocolate milk. *Journal of dairy science*.

A.A. Kuehn. 1962. Consumer brand choice as a learning process.

S. Le Cessie and JC Van Houwelingen. 1992. Ridge estimators in logistic regression. *Applied statistics*.

Chris Lederer and Sam Hill. 2001. See your brands through your customers eyes. *Harvard Business Review*, 79(6):125–133.

S.I. Lee, H. Lee, P. Abbeel, and A.Y. Ng. 2006. Efficient l1 regularized logistic regression. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.

D.C. Liu and J. Nocedal. 1989. On the limited memory bfgs method for large scale optimization. *Mathematical programming*.

D. Luo, H.G. Hosseini, and J.R. Stewart. 2004. Application of ann with extracted parameters from an electronic nose in cigarette brand identification. *Sensors and Actuators B: Chemical*.

J. C. Makens. 1965. Effect of brand preference upon consumers perceived taste of turkey meat. *Journal of Applied Psychology*.

A. F. T. Martins, N. A. Smith, P. M. Q. Aguiar, and M. A. T. Figueiredo. 2011. Structured sparsity in structured prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Y. Moon and J. Quelch. 2006. *Starbucks: delivering customer service*. Harvard Business School.

M. OConnell and G. Gooding. 2006. The use of first names to evaluate reports of gender and its effect on the distribution of married and unmarried couple households. In *Proceedings of the Annual Meetings of the Population Association of America*.

S. Ovide. 2011. Face off! dunkin' donuts vs. starbucks. In *Deal Journal - Wall Street Journal Blogs*.

M. Paul and M. Dredze. 2012. Factorial lda: Sparse multi-dimensional text models. In *Advances in Neural Information Processing Systems (NIPS)*.

F. Pelisson, D. Hall, O. Riff, and J. Crowley. 2003. Brand identification using gaussian derivative histograms. *Computer Vision Systems*.

R. Rifkin and A. Klautau. 2004. In defense of one-vs-all classification. *The Journal of Machine Learning Research (JMLR)*.

M. Schmidt, G. Fung, and R. Rosales. 2007. Fast optimization methods for l1 regularization: A comparative study and two new approaches. *Machine Learning*.

M. Schmidt, E. Van Den Berg, M. Friedlander, and K. Murphy. 2009. Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. In *Proceedings of Conference on Artificial Intelligence and Statistics (AIStats)*.

R. Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*.

W. Y. Wang and J. Hirschberg. 2011. Detecting levels of interest from spoken dialog with multistream prediction feedback and similarity based hierarchical fusion learning. In *Proceedings of the 12th annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2011)*.

W. Y. Wang, S. Finkelstein, A. Ogan, A. W. Black, and J. Cassell. 2012a. "love ya, jerkface": using sparse log-linear models to build positive (and impolite) relationships with teens. In *Proceedings of the 13th annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2012)*.

W. Y. Wang, E. Mayfield, S. Naidu, and J. Dittmar. 2012b. Historical analysis of legal opinions with a sparse mixed-effects latent variable model. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*.

K. Q. Weinberger, F. Sha, Q. Zhu, and L. K. Saul. 2007. Graph laplacian regularization for large-scale semidefinite programming. *Advances in neural information processing systems (NIPS)*.

H. Zou and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.