Modeling Social Norms Evolution for Personalized Sentiment Classification

Lin Gong[1], **Mohammad Al Boni**[2], **Hongning Wang**[1]
[1]Department of Computer Science, [2]Department of System and Information Engineering
University of Virginia, Charlottesville VA, 22904 USA
{lg5bt, ma2sm, hw5x}@virginia.edu

Abstract

Motivated by the findings in social science that people's opinions are diverse and variable while together they are shaped by evolving social norms, we perform personalized sentiment classification via *shared model adaptation* over time. In our proposed solution, a global sentiment model is constantly updated to capture the homogeneity in which users express opinions, while personalized models are simultaneously adapted from the global model to recognize the heterogeneity of opinions from individuals. Global model sharing alleviates data sparsity issue, and individualized model adaptation enables efficient online model learning. Extensive experimentations are performed on two large review collections from Amazon and Yelp, and encouraging performance gain is achieved against several state-of-the-art transfer learning and multi-task learning based sentiment classification solutions.

1 Introduction

Sentiment is personal; the same sentiment can be expressed in various ways and the same expression might carry distinct polarities across different individuals (Wiebe et al., 2005). Current mainstream solutions of sentiment analysis overlook this fact by focusing on population-level models (Liu, 2012; Pang and Lee, 2008). But the idiosyncratic and variable ways in which individuals communicate their opinions make a global sentiment classifier incompetent and consequently lead to suboptimal opinion mining results. For instance, a shared statistical classifier can hardly recognize that in restaurant reviews, the word "expensive" may indicate some users' satisfaction with a restaurant's quality, although it is generally associated with negative attitudes. Hence, a personalized sentiment classification solution is required to achieve fine-grained understanding of individuals' distinctive and dynamic opinions and benefit downstream opinion mining applications.

Sparse observations of individuals' opinionated data (Max, 2014) prevent straightforward solutions from building personalized sentiment classification models, such as estimating supervised classifiers on a per-user basis. Semi-supervised methods are developed to address the data sparsity issue. For example, leveraging auxiliary information from user-user and user-document relations in transductive learning (Hu et al., 2013; Tan et al., 2011). However, only one global model is estimated there, and the details of how individual users express diverse opinions cannot be captured. More importantly, existing solutions build static sentiment models on historic data; but the means in which a user expresses his/her opinion is changing over time. To capture temporal dynamics in a user's opinions with existing solutions, repeated model reconstruction is unavoidable, albeit it is prohibitively expensive. As a result, personalized sentiment analysis requires effective exploitation of users' own opinionated data and efficient execution of model updates across all users.

To address these challenges, we propose to build personalized sentiment classification models via *shared model adaptation*. Our solution roots in the social psychology theories about humans' dispositional tendencies (Briley et al., 2000). Humans' behaviors are shaped by social norms, a set of socially shared "feelings" and "display rules" about how one should feel and express opinions (Barsäde and Gibson, 1998; Sherif, 1936). In the context of content-based sentiment classification, we interpret social norms as global model sharing and adaptation across users. Formally, we assume a global sentiment model serves as the basis to capture self-enforcing sentimental regulari-

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 855–865,
Berlin, Germany, August 7-12, 2016.

ties across users, and each individual user tailors the shared model to realize his/her personal preference. In addition, social norms also evolve over time (Ehrlich and Levin, 2005), which leads to shifts in individuals' behaviors. This can again be interpreted as model adaptation: a new global model is adapted from an existing one to reflect the newly adopted sentimental norms. The temporal changes in individuals' opinions can be efficiently captured via online model adaptation at the levels of both global and personalized models.

Our proposed solution can also be understood from the perspective of multi-task learning (Evgeniou and Pontil, 2004; Jacob et al., 2009). Intuitively, personalized model adaptations can be considered as a set of related tasks in individual users, which contribute to a shared global model adaptation. In particular, we assume the distinct ways in which users express their opinions can be characterized by a linear classifier's parameters, i.e., the weights of textual features. Personalized models are thus achieved via a series of linear transformations over a globally shared classifier's parameters (Wang et al., 2013), e.g., shifting and scaling the weight vector. This globally shared classifier itself is obtained via another set of linear transformations over a given base classifier, which can be estimated from an isolated collection beforehand and serves as a prior for shared sentiment classification. The shared global model adaptation makes personalized model estimation no longer independent, such that regularity is formed across individualized learning tasks.

We empirically evaluated the proposed solution on two large collections of reviews, i.e., Amazon and Yelp reviews. Extensive experiment results confirm its effectiveness: the proposed method outperformed user-independent classification methods, several state-of-the-art model adaption methods, and multi-task learning algorithms.

2 Related Work

Text-based sentiment classification forms the foundation of sentiment analysis (Liu, 2012; Pang and Lee, 2008). There are two typical types of studies in sentiment classification. The first is classifying input text units (such as documents, sentences and phrases) into predefined categories, e.g., positive v.s., negative (Pang et al., 2002; Gao et al., 2014) and multiple classes (Pang and Lee, 2005). Both lexicon-based and learning-based solutions have been explored. The second is identifying topical aspects and corresponding opinions, e.g., developing topic models to predict fine-grained aspect ratings (Titov and McDonald, 2008; Wang et al., 2011). However, all those works emphasize population-level analysis, which applies a global model on all users and therefore fails to recognize the heterogeneity in which different users express their diverse opinions.

Our proposed solution is closely related to multi-task learning, which exploits the relatedness among multiple learning tasks to benefit each single task. Tasks can be related in various ways. A typical assumption is that all learnt models are close to each other in some matrix norms (Evgeniou and Pontil, 2004; Jacob et al., 2009). This has been empirically proved to be effective for capturing preferences of individual users (Evgeniou et al., 2007). Task relatedness has also been imposed via constructing a common underlying representation across different tasks (Argyriou et al., 2008; Evgeniou and Pontil, 2007). Our solution postulates task relatedness via a two-level model adaptation procedure. The global model adaptation accounts for the homogeneity and shared dynamics in users' opinions; and personalized model adaptation realizes heterogeneity in individual users.

The idea of model adaptation has been extensively explored in the context of transfer learning (Pan and Yang, 2010), which focuses on applying knowledge gained while solving one problem to different but related problems. In opinion mining community, transfer learning is mostly exploited for domain adaptation, e.g., adapting sentiment classifiers trained on book reviews to DVD reviews (Blitzer et al., 2006; Pan et al., 2010). Personalized model adaptation has also been studied in literature. The idea of linear transformation based model adaptation is introduced in (Wang et al., 2013) for personalized web search. Al Boni et al. applied a similar idea to achieve personalized sentiment classification (Al Boni et al., 2015). (Li et al., 2010) developed an online learning algorithm to continue training personalized classifiers based on a given global model. However, all of these aforementioned solutions perform model adaptation from a fixed global model, such that the learning of personalized models is independent from each other. Data sparsity again is the major bottleneck for such solutions. Our solution associates individual model adaptation via a shared global model adaptation, which leverages observations across users and thus reduces preference learning complexity.

3 Methodology

We propose to build personalized sentiment classifiers via shared model adaptation for both a global sentiment model and individualized models. Our solution roots in the social psychology theories about humans' dispositional tendencies, e.g., social norms and the evolution of social norms over time. In the following discussions, we will first briefly discuss the social theories that motivate our research, and then carefully describe the model assumptions and technical details about the proposed personalized model adaptation solution.

3.1 The Evolution of Social Norms

Social norms create pressures to establish socialization of affective experience and expression (Shott, 1979). Within the limit set by social norms and internal stimuli, individuals construct their sentiment, which is not automatic, physiological consequences but complex consequences of learning, interpretation, and social influence. This motivates us to build a global sentiment classification model to capture the shared basis on which users express their opinions. For example, the phrase "a waste of money" generally represents negative opinions across all users; and it is very unlikely that anybody would use it in a positive sense. On the other hand, members of some segments of a social structure tend to feel certain emotions more often or more intensely than members of other segments (Hochschild, 1975). Personalized model adaptation from the shared global model becomes necessary to capture the variability in affective expressions across users. For example, the word "expensive" may indicate some users' satisfaction with their received service.

Studies in social psychology also suggest that social norms shift and spread through infectious transfer mediated by webs of contact and influence over time (Ostrom, 2014; Ehrlich and Levin, 2005). Members inside a social structure influence the other members; confirmation of shifted beliefs leads to the development and evolution of social norms, which in turn regulate the shared social behaviors as a whole over time. The evolving nature of social norms urges us to take a dynamic view of the shared global sentiment model: instead of treating it as fixed, we further assume this model is also adapted from a predefined one, which serves as prior for sentiment classification. All individual users are coupled and contribute to this shared global model adaptation. This two-level model adaptation assumption leads us to the proposed multi-task learning solution, which will be carefully discussed in the next section.

3.2 Shared Linear Model Adaptation

In this paper, we focus on linear models for personalized sentiment classification due to their empirically superior performance in text-based sentiment analysis (Pang et al., 2002; Pang and Lee, 2005). We assume the diverse ways in which users express their opinions can be characterized by different settings of a linear model's parameters, i.e., the weights of textual features.

Formally, we denote a given set of opinionated text documents from user u as $D^u = \{(x_d^u, y_d^u)\}_{d=1}^{|D^u|}$, where each document x_d^u is represented by a V-dimensional vector of textual features and y_d^u is the corresponding sentiment label. The task of personalized sentiment classification is to estimate a personalized model $y = f^u(x)$ for user u, such that $f^u(x)$ best captures u's opinions in his/her generated text content. Instead of assuming $f^u(x)$ is solely estimated from user u's own opinionated data, which is prone to overfitting, we assume it is derived from a globally shared sentiment model $f^s(x)$ via model adaptation (Al Boni et al., 2015; Wang et al., 2013), i.e., shifting and scaling $f^s(x)$'s parameters for each individual user. To simplify the following discussions, we will focus on binary classification, i.e., $y_d \in \{0, 1\}$, and use the logistic regression as our reference model. But the developed techniques are general and can be easily extended to multi-class classification and generalized linear models.

We only consider scaling and shifting operations, given rotation requires to estimate much more free parameters (i.e., $O(V^2)$ v.s., $O(V)$) but contributes less in final classification performance (Al Boni et al., 2015). We further assume the adaptations can be performed in a group-wise manner (Wang et al., 2013): features in the same group will be updated synchronously by enforcing the same shifting and scaling operations. This enables the observations from seen features to be propagated to unseen features in the same group during adaptation. Various feature grouping methods have been explored in (Wang et al., 2013).

Specifically, we define $g(i) \rightarrow j$ as a feature grouping method, which maps feature i in $\{1, 2, \ldots, V\}$ to feature group j in $\{1, 2, \ldots, K\}$. A personalized model adaptation matrix can then be represented as a $2K$-dimensional vector $A^u = (a_1^u, a_2^u, \ldots, a_K^u, b_1^u, b_2^u, \ldots, b_K^u)$, where a_k^u and b_k^u

represent the scaling and shifting operations in feature group k for user u accordingly. Plugging this group-wise model adaptation into the logistic function, we can get a personalized logistic regression model $P^u(y_d = 1|x_d)$ for user u as follows,

$$P^u(y_d = 1|x_d) = \frac{1}{1 + e^{-\sum_{k=1}^{K} \sum_{g(i)=k} (a_k^u w_i^s + b_k^u) x_i}} \quad (1)$$

where w^s is the feature weight vector in the global model $f^s(x)$. As a result, personalized model adaptation boils down to identifying the optimal model transformation operation A^u for each user based on w^s and D^u.

In (Al Boni et al., 2015; Wang et al., 2013), $f^s(x)$ is assumed to be given and fixed. It leads to isolated estimation of personalized models. Based on the social norms evolution theory, $f^s(x)$ should also be dynamic and ever-changing to reflect shifted social norms. Hence, we impose another layer of model adaptation on top of the shared global sentiment model $f^s(x)$, by assuming itself is also adapted from a predefined base sentiment model. Denote this base classifier as $f^0(x)$, which is parameterized by a feature weight vector w^0 and serves as a prior for sentiment classification. Then w^s can be derived via the same aforementioned model adaptation procedure: $w^s = A^s \tilde{w}^0$, where $\tilde{w}^0$ is an augmented vector of w^0, i.e., $\tilde{w}^0 = (w^0, 1)$, to facilitate shifting operations, and A^s is the adaptation matrix for the shared global model. We should note A^s can take a different configuration (i.e., feature groupings) from individual users' adaptation matrices.

Putting these two levels of model adaptation together, a personalized sentiment classifier is achieved via,

$$w^u = A^u A^s \tilde{w}^0 \quad (2)$$

which can then be plugged into Eq (1) for personalized sentiment classification.

We name this resulting algorithm as Mutli-Task Linear Model Adaptation, or *MT-LinAdapt* in short. The benefits of shared model adaptation defined in Eq (2) are three folds. First, the homogeneity in which users express their diverse opinions are captured in the jointly estimated sentiment model $f^s(x)$ across users. Second, the learnt individual models are coupled together to reduce preference learning complexity, i.e., they collaboratively serve to reduce the models' overall prediction error. Third, non-linearity is achieved via the two-level model adaptation, which introduces more flexibility in capturing heterogeneity in different users' opinions. In-depth discussions of those unique benefits will be provided when we introduce the detailed model estimation methods.

3.3 Joint Model Estimation

The ideal personalized model adaptation should be able to adjust the individualized classifier $f^u(x)$ to minimize misclassification rate on each user's historical data in D^u. In the meanwhile, the shared sentiment model $f^s(x)$ should serve as the basis for each individual user to reduce the prediction error, i.e., capture the homogeneity. These two related objectives can be unified under a joint optimization problem.

In logistic regression, the optimal adaptation matrix A^u for an individual user u, together with A^s can be retrieved by a maximum likelihood estimator (i.e., minimizing logistic loss on a user's own opinionated data). The log-likelihood function in each individual user is defined as,

$$L(A^u, A^s) = \sum_{d=1}^{|D^u|} \Big[y_d \log P^u(y_d = 1|x_d) + (1 - y_d) \log P^u(y_d = 0|x_d) \Big] \quad (3)$$

To avoid overfitting, we penalize the transformations which increase the discrepancy between the adapted model and its source model (i.e., between w^u and w^s, and between w^s and w^0) via a L2 regularization term,

$$R(A) = \frac{\eta_1}{2} ||\boldsymbol{a} - 1||_2 + \frac{\eta_2}{2} ||\boldsymbol{b}||_2 \quad (4)$$

and it enforces scaling to be close to one and shifting to be close to zero.

By defining a new model adaptation matrix $\mathring{A} = \{A^{u_1}, A^{u_2}, \ldots, A^{u_N}, A^s\}$ to include all unknown model adaptation parameters for individual users and shared global model, we can formalize the joint optimization problem in MT-LinAdapt as,

$$\max L(\mathring{A}) = \sum_{i=1}^{N} \Big[L(A^{u_i}) - R(A^{u_i}) \Big] - R(A^s) \quad (5)$$

which can be efficiently solved by a gradient-based optimizer, such as quasi-Newton method (Zhu et al., 1997).

Direct optimization over $\mathring{A}$ requires synchronization among all the users. But in practice, users will generate their opinionated data with different paces, such that we have to postpone model adaptation until all the users have at least one observation to update their own adaptation matrix.

This delayed model update is at high risk of missing track of active users' recent opinion changes, but timely prediction of users' sentiment is always preferred. To monitor users' sentiment in realtime, we can also estimate MT-LinAdapt in an asynchronized manner: whenever there is a new observation available, we update the corresponding user's personalized model together with the shared global model immediately. i.e., online optimization of MT-LinAdapt.

This asychronized estimation of MT-LinAdapt reveals the insight of our two-level model adaptation solution: the immediate observations in user u will not only be used to update his/her own adaptation parameters in A^u, but also be utilized to update the shared global model, thus to influence the other users, who do not have adaptation data yet. Two types of competing force drive the adaptation among all the users: $w_s = A^s \tilde{w}_0$ requires timely update of global model across users; and $w_u = A^u w_s$ enforces the individual user to conform to the newly updated global model. This effect can be better understood with the actual gradients used in this asychronized update. We illustrate the decomposed gradients for scaling operation in A^u and A^s from the log-likelihood part in Eq (5) on a specific adaptation instance (x_d^u, y_d^u):

$$\frac{\partial L(A^u,A^s)}{\partial a_k^u} = \Delta_d^u \sum_{g^u(i)=k} \left(a^s_{g^s(i)} w_i^0 + b^s_{g^s(i)} \right) x^u_{di} \quad (6)$$

$$\frac{\partial L(A^u,A^s)}{\partial a_l^s} = \Delta_d^u \sum_{g^s(i)=l} a^u_{g^u(i)} w_i^0 x^u_{di} \quad (7)$$

where $\Delta_d^u = y_d^u - P^u(y_d^u = 1|x_d^u)$, and $g^u(\cdot)$ and $g^s(\cdot)$ are feature grouping functions in individual user u and shared global model $f^s(x)$.

As stated in Eq (6) and (7), the update of scaling operation in the shared global model and individual users depends on each other; the gradient with respect to global model adaptation will be accumulated among all the users. As a result, all users are coupled together via the global model adaptation in MT-LinAdapt, such that model update is propagated through users to alleviate data sparsity issue in each single user. This achieves the effect of multi-task learning. The same conclusion also applies to the shifting operations.

It is meaningful for us to compare our proposed MT-LinAdapt algorithm with those discussed in the related work section. Different from the model adaptation based personalized sentiment classification solution proposed in (Al Boni et al., 2015), which treats the global model as fixed, MT-LinAdapt adapts the global model to capture the evolving nature of social norms. As a result, in (Al Boni et al., 2015) the individualized model adaptations are independent from each other; but in MT-LinAdapt, the individual learning tasks are coupled together to enable observation sharing across tasks, i.e., multi-task learning. Additionally, as illustrated in Eq (6) and (7), nonlinear model adaptation is achieved in MT-LinAdapt because of the different feature groupings in individual users and global model. This enables observations sharing across different feature groups, while in (Al Boni et al., 2015) observations can only be shared within the same feature group, i.e., linear model adaptation. Multi-task SVM introduced in (Evgeniou and Pontil, 2004) can be considered as a special case of MT-LinAdapt. In Multi-task SVM, only shifting operation is considered in individual users and the global model is simply estimated from the pooled observations across users. Therefore, only linear model adaptation is achieved in Multi-task SVM and it cannot leverage prior knowledge conveyed in a predefined sentiment model.

4 Experiments

In this section, we perform empirical evaluations of the proposed MT-LinAdapt model. We verified the effectiveness of different feature groupings in individual users' and shared global model adaptation by comparing our solution with several state-of-the-art transfer learning and multi-task learning solutions for personalized sentiment classification, together with some qualitative studies to demonstrate how our model recognizes users' distinct expressions of sentiment.

4.1 Experiment Setup

• **Datesets.** We evaluated the proposed model on two large collections of review documents, i.e., Amazon product reviews (McAuley et al., 2015) and Yelp restaurant reviews (Yelp, 2016). Each review document contains a set of attributes such as author ID, review ID, timestamp, textual content, and an opinion rating in discrete five-star range. We applied the following pre-processing steps on both datasets: 1) filtered duplicated reviews; 2) labeled reviews with overall rating above 3 stars as positive, below 3 stars as negative, and removed the rest; 3) removed reviewers who posted more than 1,000 reviews and those whose positive review ratio is more than 90% or less than 10%

(little variance in their opinions and thus easy to classify). Since such users can be easily captured by the base model, the removal emphasizes comparisons on adapted models; 4) sorted each user's reviews in chronological order. Then, we performed feature selection by taking the union of top unigrams and bigrams ranked by Chi-square and information gain metrics (Yang and Pedersen, 1997), after removing a standard list of stopwords and porter stemming. The final controlled vocabulary consists of 5,000 and 3,071 textual features for Amazon and Yelp datasets respectively; and we adopted TF-IDF as the feature weighting scheme. From the resulting data sets, we randomly sampled 9,760 Amazon reviewers and 11,733 Yelp reviewers for testing purpose. There are 105,472 positive reviews and 37,674 negative reviews in the selected Amazon dataset; 108,105 positive reviews and 32,352 negative reviews in the selected Yelp dataset.

• **Baselines.** We compared the performance of MT-LinAdapt against seven different baselines, ranging from user-independent classifiers to several state-of-the-art model adaption methods and multi-task learning algorithms. Due to space limit, we will briefly discuss the baseline models below.

Our solution requires a user-independent classifier as base sentiment model for adaptation. We estimated logistic regression models from a separated collection of reviewers outside the preserved testing data on Amazon and Yelp datasets accordingly. We also included these isolated base models in our comparison and name them as *Base*. In order to verify the necessity of personalized sentiment models, we trained a global SVM based on the pooled adaptation data from all testing reviewers, and name it as *Global SVM*. We also estimated an independent SVM model for each single user only based on his/her adaptation reviews, and name it as *Individual SVM*. We included an instance-based transfer learning method (Brighton and Mellish, 2002), which considers the k-nearest neighbors of each testing review document from the isolated training set for personalized model training. As a result, for each testing case, we estimated an independent classification model, which is denoted as *ReTrain*. (Geng et al., 2012) used L2 regularization to enforce the adapted models to be close to the global model. We applied this method to get personalized logistic regression models and refer to it as *RegLR*. *LinAdapt* developed in (Al Boni et al., 2015) also performs group-wise linear model adaptation to build personalization classifiers. But it isolates model adaptation in individual users. *MT-SVM* is a multi-task learning method, which encodes task relatedness via a shared linear kernel (Evgeniou and Pontil, 2004).

• **Evaluation Settings.** We evaluated all the models with both synchronized (batch) and asynchronized (online) model update. We should note MT-SVM can only be tested in batch mode, because it is prohibitively expensive to retrain SVM repeatedly. In batch evaluation, we split each user's reviews into two sets: the first 50% for adaptation and the rest 50% for testing. In online evaluation, once we get a new testing instance, we first evaluate the up-to-date personalized classifier against the ground-truth; then use the instance to update the personalized model. To simulate the real-world situation where user reviews arrive sequentially and asynchronously, we ordered all reviews chronologically and accessed them one at a time for online model update. In particular, we utilized stochastic gradient descent for this online optimization (Kiwiel, 2001). Because of the biased class distribution in both datasets, we computed F1 measure for both positive and negative class in each user, and took macro average among users to compare the different models' performance.

4.2 Effect of Feature Grouping

In MT-LinAdapt, different feature groupings can be postulated in individual users' and shared global model adaptation. Nonlinearity is introduced when different grouping functions are used in these two levels of model adaptation. Therefore, we first investigated the effect of feature grouping in MT-LinAdapt.

We adopted the feature grouping method named "*cross*" in (Wang et al., 2013) to cluster features into different groups. More specifically, we evenly spilt the training collection into N non-overlapping folds, and train a single SVM model on each fold. Then, we create a $V \times N$ matrix by putting the learned weights from N folds together, on which k-means clustering is applied to extract K feature groups. We compared the batch evaluation performance of varied combinations of feature groups in MT-LinAdapt. The experiment results are demonstrated in Table 1; and for comparison purpose, we also included the base classifier's performance in the table.

In Table 1, the two numbers in the first column denote the feature group sizes in personalized models and global model respectively. And *all* indicates one feature per group (i.e., no fea-

Table 1: Effect of different feature groupings in MT-LinAdapt.

Method	Amazon		Yelp	
	Pos F1	Neg F1	Pos F1	Neg F1
Base	0.8092	0.4871	0.7048	0.3495
400-800	0.8318	0.5047	0.8237	0.4807
400-1600	0.8385	0.5257	0.8309	0.4978
400-all	0.8441	0.5423	0.8345	0.5105
800-800	0.8335	0.5053	0.8245	0.4818
800-1600	0.8386	0.5250	0.8302	0.4962
800-all	0.8443	**0.5426**	**0.8361**	**0.5122**
1600-all	**0.8445**	0.5424	0.8357	0.5106
all-all	0.8438	0.5416	0.8361	0.5100

ture grouping). The adapted models in MT-LinAdapt achieved promising performance improvement against the base sentiment classifier, especially on the Yelp data set. As we increased the feature group size for global model, MT-LinAdapt's performance kept improving; while with the same feature grouping in the shared global model, a moderate size of feature groups in individual users is more advantageous.

These observations are expected. Because the global model is shared across users, all their adaptation reviews can be leveraged to adapt the global model so that sparsity is no longer an issue. Since more feature groups in the global model can be afforded, more accurate estimation of adaptation parameters can be achieved. But at the individual user level, data sparsity is still the bottleneck for accurate adaptation estimation, and trade-off between observation sharing and estimation accuracy has to be made. Based on this analysis, we selected **800** and **all** feature groups for individual models and global model respectively in the following experiments.

4.3 Personalized Sentiment Classification

• **Synchronized model update.** Table 2 demonstrated the classification performance of MT-LinAdapt against all baselines on both Amazon and Yelp datasets, where binomial tests on win-loss comparison over individual users were performed between the best algorithm and runner-up to verify the significance of performance improvement. We can clearly notice that MT-LinAdapt significantly outperformed all baselines in negative class, and it was only slightly worse than MT-SVM on positive class. More specifically, per-user classifier estimation clearly failed to obtain a usable classifier, due to the sparse observations in single users. Model-adaptation based baselines, i.e., RegLR and LinAdapt, slightly improved over the base model. But because the adaptations across users are isolated and the base model is fixed, their improvement is very limited. As for negative class, MT-LinAdapt outperformed Global SVM significantly on both datesets. Since negative class suffers more from the biased prior distribution, the considerable performance improvement indicates effectiveness of our proposed personalized sentiment classification solution. As for positive class, the performance difference is not significant between MT-LinAdapt and MT-SVM on Amazon data set nor between MT-LinAdapt and Global SVM on Yelp data set. By looking into detailed results, we found that MT-LinAdapt outperformed MT-SVM on users with fewer adaptation reviews. Furthermore, though MT-SVM benefits from multi-task learning, it cannot leverage information from the given base classifier. Considering the biased class prior in these two data sets (2.8:1 on Amazon and 3.3:1 on Yelp), the improved classification performance on negative class from MT-LinAdapt is more encouraging.

Table 2: Classification results in batch mode.

Method	Amazon		Yelp	
	Pos F1	Neg F1	Pos F1	Neg F1
Base	0.8092	0.4871	0.7048	0.3495
Global SVM	0.8352	0.5403	**0.8411**	0.5007
Individual SVM	0.5582	0.2418	0.3515	0.3547
ReTrain	0.7843	0.4263	0.7807	0.3729
RegLR	0.8094	0.4896	0.7103	0.3566
LinAdapt	0.8091	0.4894	0.7107	0.3575
MT-SVM	**0.8484**	0.5367	0.8408	0.5079
MT-LinAdapt	0.8441	**0.5422***	0.8358	**0.5119***

$*$ indicates p-value$<$0.05 with Binomial test.

• **Asynchronized model update.** In online model estimation, classifiers can benefit from immediate update, which provides a feasible solution for timely sentiment analysis in large datasets. In this setting, only two baseline models are applicable without model reconstruction, i.e., RegLR and LinAdapt. To demonstrate the utility of online update in personalized sentiment models, we illustrate the relative performance gain of these models over the base sentiment model in Figure 1. The x-axis indicates the number of adaptation instances consumed in online update from all users, i.e., the 1st review means after collecting the first review of each user.

MT-LinAdapt converged to satisfactory performance with only a handful of observations in each user. LinAdapt also quickly converged, but its performance was very close to the base model, since no observation is shared across users. RegLR needs the most observations to estimate satisfac-

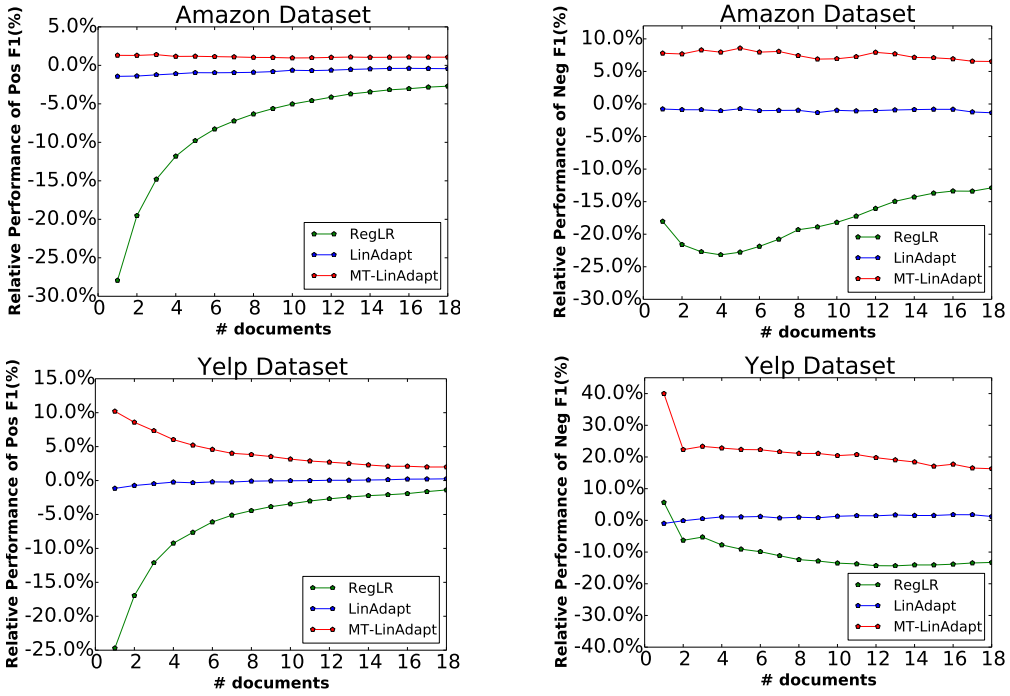

Figure 1: Relative performance gain between MT-LinAdapt and baselines on Amazon and Yelp datasets.

tory personalized models. The improvement in MT-LinAdapt demonstrates the benefit of shared model adaptation, which is vital when the individuals' adaptation data are not immediately available but timely sentiment classification is required.

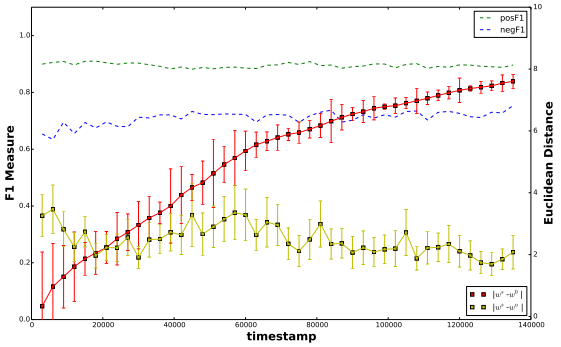

Figure 2: Online model update trace on Amazon.

It is meaningful to investigate how the shared global model and personalized models are updated during online learning. The shift in the shared global model reflects changes in social norms, and the discrepancy between the shared global model and personalized models indicates the variances of individuals' opinions. In particular, we calculated Euclidean distance between global model w^s and base model w^0 and that between individualized model w^u and shared global model w^s during online model updating. To visualize the results, we computed and plotted the average Euclidean distances in every 3000 observations during online learning, together with the corresponding variance. To illustrate a comprehensive picture of online model update, we also plotted the corresponding average F1 performance for both positive and negative class. Because the Euclidean distance between w^s and w^0 is much larger than that between w^s and w^u, we scaled $||w^s - w^0||$ by 0.02 on Amazon dataset in Figure 2. Similar results were observed on Yelp data as well; but due to space limit, we do not include them.

As we can clearly observe that the difference between the base model and newly adapted global model kept increasing during online update. At the earlier stage, it is increasing much faster than the later stage, and the corresponding classification performance improves more rapidly (especially in negative class). The considerably large variance between w^0 and w^s at the beginning indicates the divergence between old and new social norms across users. Later on, variance decreased and converged with more observations, which can be understood as the formation of the new social norms among users. On the other hand, the distance between personalized models and shared global model fluctuated a lot at the beginning; with more observations, it became stable later on. This is also reflected in the range of variance: the variance is much smaller in later stage than earlier stage, which indicates users comply to the newly established social norms.

Table 3: Shared model adaptation for cold start on Amazon and Yelp.

Obs.	Amazon						Yelp					
	Shared-SVM		MT-SVM		MT-LinAdapt		Shared-SVM		MT-SVM		MT-LinAdapt	
	Pos F1	Neg F1	Pos F1	Neg F1	Pos F1	Neg F1	Pos F1	Neg F1	Pos F1	Neg F1	Pos F1	Neg F1
1^{st}	0.9004	0.7013	**0.9264**	0.7489	0.9122	**0.7598**	0.7882	0.5537	**0.9040**	0.7201	0.8809	**0.7306**
2^{nd}	0.9200	0.6872	**0.9200**	**0.7319**	0.8945	0.7292	0.7702	0.5266	**0.8962**	0.6959	0.8598	**0.6968**
3^{rd}	0.9164	0.6967	**0.9164**	0.7144	0.8967	**0.7260**	0.7868	0.5278	**0.9063**	**0.7099**	0.8708	0.7069

4.4 Shared Adaptation Against Cold Start

Cold start refers to the challenge that a statistic model cannot draw any inference for users before sufficient observations are gathered (Schein et al., 2002). The shared model adaptation in MT-LinAdapt helps alleviate cold start in personalized sentiment analysis, while individualized model adaptation method, such as RegLR and LinAdapt, cannot achieve so. To verify this aspect, we separated both Amazon and Yelp reviewers into two sets: we randomly selected 1,000 reviewers from the isolated training set and exhausted all their reviews to estimate a shared SVM model, MT-LinAdapt and MT-SVM. Then those models were directly applied onto the testing reviewers for evaluation. Again, because it is time consuming to retrain a SVM model repeatedly, only MT-LinAdapt performed online model update in this evaluation. We report the performance on the first three observations from all testing users accordingly in Table 3.

MT-LinAdapt achieved promising performance on the first testing cases, especially on the negative class. This indicates its estimated global model is more accurate on the new testing users. Because MT-SVM cannot be updated during this online test, only its previously estimated global model from the 1,000 training users can be applied here. As we can notice, its performance is very similar to the shared SVM model (especially on Amazon dataset). MT-LinAdapt adapts to this new collection of users very quickly, so that improved performance against the static models at later stage is achieved.

4.5 Vocabulary Stability

One derivative motivation for personalized sentiment analysis is to study the diverse use of vocabulary across individual users. We analyzed the variance of words' sentiment polarities estimated in the personalized models against the base model. Table 4 shows the most and the least variable features on both datasets. It is interesting to find that words with strong sentiment polarities tend to be more stable across users, such as "disgust," "regret," and "excel." This demonstrates the sign of conformation to social norms. There are also words exhibiting high variances in sentiment polarity, such as "was-yummi," "lazi," and "cheat," which indicates the heterogeneity of users' opinionated expressions.

Table 4: Top six words with the highest and lowest variances of learned polarities by MT-LinAdapt.

Amazon	Highest	cheat astound	healthi the-wrong	enjoy-read the-amaz
	Lowest	mistak regret	favor perfect-for	excel great
Yelp	Highest	total-worth advis	lazi impress	was-yummi so-friend
	Lowest	omg frustrat	veri-good disgust	hungri a-must

5 Conclusions

In this work, we proposed to perform personalized sentiment classification based on the notion of shared model adaptation, which is motivated by the social theories that humans' opinions are diverse but shaped by the ever-changing social norms. In the proposed MT-LinAdapt algorithm, global model sharing alleviates data sparsity issue, and individualized model adaptation captures the heterogeneity in humans' sentiments and enables efficient online model learning. Extensive experiments on two large review collections from Amazon and Yelp confirmed the effectiveness of our proposed solution.

The idea of shared model adaptation is general and can be further extended. We currently used a two-level model adaptation scheme. The adaptation can be performed at the user group level, i.e., three-level model adaptation. The user groups can be automatically identified to maximize the effectiveness of shared model adaptation. In addition, this method can also be applied to domain adaptation, where a domain taxonomy enables a hierarchically shared model adaptation.

6 Acknowledgments

We thank the anonymous reviewers for their insightful comments. This paper is based upon work supported by the National Science Foundation under grant IIS-1553568.

References

[Al Boni et al.2015] Mohammad Al Boni, Keira Qi Zhou, Hongning Wang, and Matthew S Gerber. 2015. Model adaptation for personalized opinion analysis. *In Proceedings of ACL*.

[Argyriou et al.2008] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. 2008. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272.

[Barsäde and Gibson1998] Sigal G Barsäde and Donald E Gibson. 1998. Group emotion: A view from top and bottom. *Research on managing groups and teams*, 1:81–102.

[Blitzer et al.2006] John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 EMNLP*, pages 120–128. ACL.

[Brighton and Mellish2002] Henry Brighton and Chris Mellish. 2002. Advances in instance selection for instance-based learning algorithms. *Data mining and knowledge discovery*, 6(2):153–172.

[Briley et al.2000] Donnel A Briley, Michael W Morris, and Itamar Simonson. 2000. Reasons as carriers of culture: Dynamic versus dispositional models of cultural influence on decision making. *Journal of consumer research*, 27(2):157–178.

[Ehrlich and Levin2005] Paul R Ehrlich and Simon A Levin. 2005. The evolution of norms. *PLoS Biol*, 3(6):e194.

[Evgeniou and Pontil2004] Theodoros Evgeniou and Massimiliano Pontil. 2004. Regularized multi–task learning. In *Proceedings of the 10th ACM SIGKDD*, pages 109–117. ACM.

[Evgeniou and Pontil2007] A Evgeniou and Massimiliano Pontil. 2007. Multi-task feature learning. *Advances in neural information processing systems*, 19:41.

[Evgeniou et al.2007] Theodoros Evgeniou, Massimiliano Pontil, and Olivier Toubia. 2007. A convex optimization approach to modeling consumer heterogeneity in conjoint estimation. *Marketing Science*, 26(6):805–818.

[Gao et al.2014] Wenliang Gao, Nobuhiro Kaji, Naoki Yoshinaga, and Masaru Kitsuregawa. 2014. Collective sentiment classification based on user leniency and product popularity. *Ił*, 21(3):541–561.

[Geng et al.2012] Bo Geng, Yichen Yang, Chao Xu, and Xian-Sheng Hua. 2012. Ranking model adaptation for domain-specific search. *IEEE Transactions on Knowledge and Data Engineering*, 24(4):745–758.

[Hochschild1975] Arlie Russell Hochschild. 1975. The sociology of feeling and emotion: Selected possibilities. *Sociological Inquiry*, 45(2-3):280–307.

[Hu et al.2013] Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu. 2013. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the 6th WSDM*, pages 537–546. ACM.

[Jacob et al.2009] Laurent Jacob, Jean-philippe Vert, and Francis R Bach. 2009. Clustered multi-task learning: A convex formulation. In *NIPS*, pages 745–752.

[Kiwiel2001] Krzysztof C Kiwiel. 2001. Convergence and efficiency of subgradient methods for quasiconvex minimization. *Mathematical programming*, 90(1):1–25.

[Li et al.2010] Guangxia Li, Steven CH Hoi, Kuiyu Chang, and Ramesh Jain. 2010. Micro-blogging sentiment detection by collaborative online learning. In *ICDM*, pages 893–898. IEEE.

[Liu2012] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

[Max2014] Woolf Max. 2014. A statistical analysis of 1.2 million amazon reviews. `http://minimaxir.com/2014/06/reviewing-reviews`.

[McAuley et al.2015] Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

[Ostrom2014] Elinor Ostrom. 2014. Collective action and the evolution of social norms. *Journal of Natural Resources Policy Research*, 6(4):235–252.

[Pan and Yang2010] Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359.

[Pan et al.2010] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th WWW*, pages 751–760. ACM.

[Pang and Lee2005] Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd ACL*, pages 115–124. ACL.

[Pang and Lee2008] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

[Pang et al.2002] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86. ACL.

[Schein et al.2002] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. 2002. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260. ACM.

[Sherif1936] Muzafer Sherif. 1936. The psychology of social norms.

[Shott1979] Susan Shott. 1979. Emotion and social life: A symbolic interactionist analysis. *American journal of Sociology*, pages 1317–1334.

[Tan et al.2011] Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD*, pages 1397–1405. ACM.

[Titov and McDonald2008] Ivan Titov and Ryan T McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *ACL*, volume 8, pages 308–316. Citeseer.

[Wang et al.2011] Hongning Wang, Yue Lu, and ChengXiang Zhai. 2011. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD*, pages 618–626. ACM.

[Wang et al.2013] Hongning Wang, Xiaodong He, Ming-Wei Chang, Yang Song, Ryen W White, and Wei Chu. 2013. Personalized ranking model adaptation for web search. In *Proceedings of the 36th ACM SIGIR*, pages 323–332. ACM.

[Wiebe et al.2005] Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

[Yang and Pedersen1997] Yiming Yang and Jan O Pedersen. 1997. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420.

[Yelp2016] Yelp. 2016. Yelp dataset challenge. `https://www.yelp.com/dataset_challenge`.

[Zhu et al.1997] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. 1997. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560.