Deceptive Review Spam Detection via Exploiting Task Relatedness and Unlabeled Data

Zhen Hai[†] **Peilin Zhao**[‡] **Peng Cheng**[§] **Peng Yang**[*] **Xiao-Li Li** [†] **Guangxia Li**[¶]
[†]Institute for Infocomm Research, A*STAR, Singapore, {*haiz,xlli*}*@i2r.a-star.edu.sg*
[‡]Ant Financial, Hangzhou, China, *peilin.zpl@alipay.com*
[§]SCSE, Nanyang Technological University, Singapore, *pcheng1@ntu.edu.sg*
[*]Tencent AI Lab, Shenzhen, China, *henryppyang@tencent.com*
[¶]SCST, Xidian University, Xi'an, China, *gxli@xidian.edu.cn*

Abstract

Existing work on detecting deceptive reviews primarily focuses on feature engineering and applies off-the-shelf supervised classification algorithms to the problem. Then, one real challenge would be to manually recognize plentiful ground truth spam review data for model building, which is rather difficult and often requires domain expertise in practice. In this paper, we propose to exploit the relatedness of multiple review spam detection tasks and readily available unlabeled data to address the scarcity of labeled opinion spam data. We first develop a multi-task learning method based on logistic regression (MTL-LR), which can boost the learning for a task by sharing the knowledge contained in the training signals of other related tasks. To leverage the unlabeled data, we introduce a graph Laplacian regularizer into each base model. We then propose a novel semi-supervised multi-task learning method via Laplacian regularized logistic regression (SMTL-LLR) to further improve the review spam detection performance. We also develop a stochastic alternating method to cope with the optimization for SMTL-LLR. Experimental results on real-world review data demonstrate the benefit of SMTL-LLR over several well-established baseline methods.

1 Introduction

Nowadays, more and more individuals and organizations have become accustomed to consulting user-generated reviews before making purchases or online bookings. Considering great commercial benefits, merchants, however, have tried to hire people to write undeserving positive reviews to promote their own products or services, and meanwhile to post malicious negative reviews to defame those of their competitors. The fictitious reviews and opinions, which are deliberately created in order to promote or demote targeted entities, are known as *deceptive opinion spam* (Jindal and Liu, 2008; Ott et al., 2011).

By formulating deceptive opinion spam detection as a classification problem, existing work primarily focuses on extracting different types of features and applies off-the-shelf supervised classification algorithms to the problem (Jindal and Liu, 2008; Ott et al., 2011; Feng et al., 2012; Chen and Chen, 2015). Then, one weakness of previous work lies in the demand of manually recognizing a large amount of ground truth review spam data for model training. Unlike other forms of spamming activities, such as email or web spam, deceptive opinion spam, which has been deliberately written to sound authentic, is more difficult to be recognized by manual read. In an experiment, three undergraduate students were (randomly) invited to identify spam reviews from nonspam ones in hotel domain. As shown in Table 1, their average accuracy is merely 57.3% (Ott et al., 2011). Then, given a limited set of labeled review data for a domain, e.g., hotel, it is almost impossible to build a robust classification model for detecting deceptive spam reviews in reality.

In this work, we deal with the problem of detecting a textual review as *spam* or not, i.e., *nonspam*. We consider each deceptive review spam detection problem within each domain, e.g., detecting

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1817–1826,
Austin, Texas, November 1-5, 2016.

	Judge-1	Judge-2	Judge-3
Accuracy	61.9%	56.9%	53.1%
F-spam	48.7%	30.3%	43.6%
F-nonspam	69.7%	68.8%	59.9%

Table 1: Performance of human judges for review spam detection in hotel domain (Ott et al., 2011), where F-spam/F-nonspam means F-score for spam/nonspam label.

spam hotel/restuarnt reviews from hotel/restaurnat domain, to be a different task. Previous studies have empirically shown that learning multiple related tasks simultaneously can significantly improve performance relative to learning each task independently, especially when only a few labeled data per task are available (Caruana, 1997; Bakker and Heskes, 2003; Argyriou et al., 2006). Thus, given the limited labeled review data for each domain, we formulate the review spam detection tasks for multiple domains, e.g., hotel, restaurant, and so on, as a multi-task learning problem.

We develop a multi-task learning method via logistic regression (MTL-LR) to address the problem. One key advantage of the method is that it allows to boost the learning for one review spam detection task by leveraging the knowledge contained in the training signals of other related tasks. Then, there is often a large quantity of review data freely available online. In order to leverage the unlabeled data, we introduce a graph Laplacian regularizer into each base logistic regression model. We extend MTL-LR, and propose a novel semi-supervised multi-task learning model via Laplacian regularized logistic regression (SMTL-LLR) to further boost the review spam detection performance under the multi-task learning setting. Moreover, to cope with the optimization problem for SMTL-LLR, we also develop a stochastic alternating optimization method, which is computationally efficient.

To the best of our knowledge, this is the first work that generalizes opinion spam detection from independent single-task learning to symmetric multi-task learning setting. By symmetric, we mean that the setting seeks to improve the performance of all learning tasks simultaneously. In this sense, it is different from transfer learning (Pan and Yang, 2010), where the objective is to improve the performance of a target task using information from source tasks. Under this new setting, we can exploit the commonality shared by related review spam detection tasks as well as readily available unlabeled data, and then alleviate the scarcity of labeled spam review data. Experimental results on real-world review data demonstrate the superiority of SMTL-LLR over several representative baseline methods.

The rest of this paper is organized as follows. Section 2 presents related work. Section 3 introduces the proposed methods and stochastic alternating optimization algorithm. Then, in Section 4, we present the experimental results in detail, and conclude this paper in Section 5.

2 Related Work

Previous work typically formulates deceptive opinion spam detection as a classification problem, and then presents different types of features to train supervised classification algorithms for the problem. Jindal and Liu (2008) first studied opinion spam detection problem. They built the ground truth review data set by treating the duplicate reviews in a given corpus as spam reviews and the rest as nonspam reviews. They presented review, product, and reviewer related features, and then trained logistic regression (LR) model on the features for finding fake review spam. Ott et al. (2011) created the ground truth review data via a crowd-sourcing service called Amazon Mechanical Turk[1]. They presented three different types of features for opinion spam detection, i.e., genre identification features, psycholinguistic deception features, and standard n-gram text features. They found that the supervised support vector machines (SVM) trained on the textual n-gram features can achieve good performance. Feng et al. (2012) presented syntactic stylometry features and trained SVM model for deception detection, while Chen and Chen (2015) built the SVM classifier on a diversity of features, such as content and thread features, for opinion spam detection in web forum. In addition, Li et al. (2014) employed a feature-based additive model to explore the general rule for deceptive opinion spam detection. Generally, in order to build robust supervised review spam detection models, we have to manually recognize large-scale ground truth spam data. But this could be very ex-

[1]https://www.mturk.com

pensive, and often requires domain expertise.

Though a large amount of unlabeled review data are freely available online, very limited work has been done on developing semi-supervised methods for review spam detection. Li et al. (2011) used a two-view co-training method (Blum and Mitchell, 1998) for semi-supervised learning to identify fake review spam. One limitation of the work is that it needs additional reviewer information when building model. Given a corpus of textual reviews, the reviewer related view may not be always available in reality. Moreover, the co-training method is not intrinsically geared to learning from the unlabeled review data, instead, simply makes use of the unlabeled reviews within a fully supervised learning framework, negating the semi-supervised learning benefit. For some particular scenarios, the available training data could be only a partially labeled set of positive examples, e.g., spam reviews, and a large set of unlabeled reviews. Positive unlabeled learning (PU) (De Comite et al., 1999; Liu et al., 2002) may be then used for deceptive review spam detection (Hernandez et al., 2013). However, this clearly contrasts with our problem, where our training data contains a complete labeled set of positive (spam) and negative (nonspam) reviews besides the unlabeled set of review data.

In addition, instead of detecting spam reviews directly, considerable efforts have been made to recognize review spammers, i.e., online users who have written spam reviews. Lim et al. (2010) studied different types of spamming behavioral indicators, and then used a regression method on the indicators for finding review spammers. Wang et al. (2012) investigated the relationships among reviewers, reviews, and stores, and developed a social review graph based method to identify online store spammers. Mukherjee et al. (2013) developed an author spamicity Bayesian model to exploit the observed behavioral footprints for spammer detection. In reality, a group of online users may work together to create spam reviews. Mukherjee et al. (2012) developed a group spam ranking algorithm to detect spammer groups.

Multi-task learning is a learning paradigm that seeks to boost generalization performance by learning a task together with other tasks at the same time while using a shared representation (Caruana, 1997). Most majority of existing work on multi-task learning does not infer actual task relations from training data automatically, instead, they typically make the assumptions that the relations are existent or are given as prior knowledge (Thrun and O'Sullivan, 1996; Bakker and Heskes, 2003; Evgeniou and Pontil, 2004; Argyriou et al., 2006; Liu et al., 2009). To better fit the multi-task learning model to real-world data, Zhang and Yeung (2010) proposed a convex regularization formulation named multi-task relation learning (MTRL), which can learn real relationships between tasks under a multi-task learning framework.

In this work, we focus on detecting online deceptive review spam. We formulate review spam detection for multiple domains (e.g., hotel and restaurant) as a multi-task learning problem. Following the convex framework of MTRL, we first develop a multi-task learning method via logistic regression (MTL-LR). We employ logistic regression as base classification model, because: 1) It is a robust model that does not have configuration parameters to tune; 2) It can be straightforwardly extended, and be efficiently trained using convex optimization techniques (Hoi et al., 2006; Minka, 2003); and 3) It has been shown effective for large-scale text classification and fake review detection problems (Hoi et al., 2006; Jindal and Liu, 2008). Then, to leverage the large volume of unlabeled review data, we extend the base logistic regression model, and incorporate a graph Laplacian regularizer into it. We thus develop a new semi-supervised multi-task learning paradigm via Laplacian regularized logistic regression, which is able to further boost the performance for review spam detection.

3 Methodology

3.1 Multi-task Learning via Logistic Regression

Given m review domains $\mathcal{D}_1, \ldots, \mathcal{D}_m$, we accordingly have m review spam detection tasks $\mathcal{T}_1, \ldots, \mathcal{T}_m$, which share a common feature space with d dimensions. For the task $\mathcal{T}_i$ in the domain $\mathcal{D}_i$, there is a small labeled set of l_i review examples $\mathcal{L}_i = \{(\mathbf{x}_1^i, y_1^i), \ldots, (\mathbf{x}_{l_i}^i, y_{l_i}^i)\}$, where $\mathbf{x}_j^i \in \mathbb{R}^d$ is the vectorial representation of the review j in the labeled set $\mathcal{L}_i$, and $y_j^i \in \{+1, -1\}$ refers to the *spam*

(+1) or *nonspam* (−1) label of the review.

When there is only one review spam detection task, for example, $\mathcal{T}_i$, we can use logistic regression (LR) model to learn a supervised classifier based on the labeled set $\mathcal{L}_i$. The objective function of LR for single-task learning is

$$P^i_{LR}(\mathbf{w}_i) = \frac{1}{l_i}\sum_{j=1}^{l_i}\ln(1+\exp(-y^i_j\mathbf{w}_i^\top\mathbf{x}^i_j)) + \frac{\lambda}{2}\|\mathbf{w}_i\|^2,$$

where $\mathbf{w}_i \in \mathbb{R}^d$, $\lambda > 0$ refers to regularization parameter.

Once the model is learned from solving the optimization problem, given a test review instance $\mathbf{x}_{j'}$ of the task $\mathcal{T}_i$, we can employ the model to predict it as *spam*, i.e., $\hat{y}_{j'} = 1$, with probability

$$Prob(\hat{y}_{j'} = 1) = \frac{1}{1+\exp(-\mathbf{w}_i^\top\mathbf{x}^i_{j'})}.$$

Now we have m review spam detection tasks for multiple domains, and we would learn m supervised classification models simultaneously. To achieve this, we introduce a covariance matrix Ω to represent the correlations among the m review spam detection tasks, where Ω_{ij} refers to the relation/covariance between a pair of tasks $\mathcal{T}_i$ and $\mathcal{T}_j$. Since Ω is a task covariance matrix, we require it to satisfy the constraint $\Omega \succeq 0$, i.e., positive semidefinite. We also restrict $Tr(\Omega) = 1$ without of loss of generality, since for any covariance matrix $Tr(\Sigma) \neq 1$, we can use $\frac{\Sigma}{Tr(\Sigma)}$ as Ω. If the covariance matrix is given as prior knowledge, then we introduce a supervised multi-task learning (MTL) framework via logistic regression as follows

$$P^\Omega_{MTL}(\mathbf{W}) = \sum_{i=1}^{m}\frac{1}{l_i}\sum_{j=1}^{l_i}\ln(1+\exp(-y^i_j\mathbf{w}_i^T\mathbf{x}^i_j)) + \frac{\lambda}{2}Tr(\mathbf{W}\mathbf{W}^\mathbf{T}) + \frac{\beta}{2}Tr(\mathbf{W}\mathbf{\Omega}^{-1}\mathbf{W}^T),$$

where $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_m)$, and $\beta > 0$ is a regularization parameter.

Under this multi-task learning setting, the first term refers to the sum of all the average empirical loss, the second term refers to the regularizer used to avoid over-fitting, and the last term is introduced to leverage the shared knowledge from multiple learning tasks according to their relationships.

In reality, the covariance matrix may be not provided a priori. We then present the following multi-task learning model, which can learn the model parameters W and Ω automatically from training review data

$$P_{MTL}(\mathbf{W}, \mathbf{\Omega}) = \sum_{i=1}^{m}\frac{1}{l_i}\sum_{j=1}^{l_i}\ln(1+\exp(-y^i_j\mathbf{w}_i^T\mathbf{x}^i_j)) + \frac{\lambda}{2}Tr(\mathbf{W}\mathbf{W}^\mathbf{T}) + \frac{\beta}{2}Tr(\mathbf{W}\mathbf{\Omega}^{-1}\mathbf{W}^T)$$
$$s.t. \quad \mathbf{\Omega} \succeq 0, \quad Tr(\mathbf{\Omega}) = 1,$$

If we have only one review spam detection task, i.e., $m = 1$, then it is straightforward to verify that the above multi-task learning formulation would be reduced to single-task objective function of logistic regression.

3.2 Semi-supervised Multi-task Learning via Laplacian Regularized Logistic Regression

Generally, for a given review domain $\mathcal{D}_i$, there is a large set of unlabeled reviews $\mathcal{U}_i = \{\mathbf{x}^i_{l_i+1}, \ldots, \mathbf{x}^i_{n_i}\}$ in addition to the labeled review set $\mathcal{L}_i$. Then, for each review spam detection task $\mathcal{T}_i$, we constracut a weighted neighborhood graph $\mathcal{G}_i = (V_i, E_i)$ based on both labeled and unlabeled review sets $\mathcal{L}_i$ and $\mathcal{U}_i$. V refers to the set of data points, each of which stands for a review example $\mathbf{x}^i_j$ ($j : 1, \ldots, n_i$) from either $\mathcal{L}_i$ or $\mathcal{U}_i$. E refers to the set of weighted edges. Specifically, if a review example/point $\mathbf{x}^i_j$ is among the K nearest neighbors of the review point $\mathbf{x}^i_k$, we put an edge linking the two examples, and vice versa. We also assign an adjacent weight score s^i_{jk} to the edge, which represents the similarity or closeness between the two reviews. Once the neighborhood graph $\mathcal{G}_i$ has been built for each task, a Laplacian regularizer can be then constrcted on the graph to extend the regular logistic regression model.

Considering the similarity matrix S_i that corresponds to the graph $\mathcal{G}_i$ for the task $\mathcal{T}_i$, it is expected that a good model would also minimize the follow-

ing objective

$$\sum_{jk} s^i_{jk}(\mathbf{w}_i^\top \mathbf{x}^i_j - \mathbf{w}_i^\top \mathbf{x}^i_k)^2,$$

This objective implies that $\mathbf{w}_i^\top \mathbf{x}^i_j$ should be close to $\mathbf{w}_i^\top \mathbf{x}^i_k$ if the similarity s^i_{jk} is large. The objective can be simplified as

$$\begin{aligned}&\sum_{jk} s^i_{jk}(\mathbf{w}_i^\top \mathbf{x}^i_j - \mathbf{w}_i^\top \mathbf{x}^i_k)^2\\ &= \mathrm{Tr}(\mathbf{w}_i^\top X_i(D_i - S_i)X_i^\top \mathbf{w}_i)\\ &= \mathrm{Tr}(\mathbf{w}_i^\top X_i L_i X_i^\top \mathbf{w}_i),\end{aligned}$$

where $D_i = diag(D^i_{jj})$ is a diagonal matrix, $D^i_{jj} = \sum_k s^i_{jk}$, and $L_i = D_i - S_i$ refers to the graph Laplacian matrix.

Then, given both labeled review set $\mathcal{L}_i$ and unlabeled set $\mathcal{U}_i$ for the task T_i, we extend the basic logistic regression by incorporating the graph Laplacian regularizer into its learning framework, and develop a new semi-supervised Laplacian regularized logistic regression (LLR) model. The objective function of LLR for semi-supervised single-task learning is given below

$$\begin{aligned}&P^i_{LLR}(\mathbf{w}_i)\\ &= \frac{1}{l_i}\sum_{j=1}^{l_i} \ln(1+\exp(-y^i_j \mathbf{w}_i^\top \mathbf{x}^i_j))\\ &+\frac{\lambda}{2}\|\mathbf{w}_i\|^2 + \frac{\gamma}{2}\mathrm{Tr}(\mathbf{w}_i^\top X_i L_i X_i^\top \mathbf{w}_i),\end{aligned}$$

where $\lambda > 0$ and $\gamma > 0$ are regularization parameters.

The semi-supervised formulation of LLR balances several desires. The first term is used to minimize the loss of the model on the labeled review data, the second term is used to minimize the complexity of the model, and the last term refers to the Laplacian regularizer, which is introduced to make the prediction of the model smooth on the whole review data set.

Next, based on the objective function of the above LLR model, we extend the supervised multi-task learning framework, and propose a novel semi-supervised multi-task learning paradigm via Laplacian regularized logistic regression (SMTL-LLR) as follows

$$\begin{aligned}&P_{SMTL}(\mathbf{W}, \mathbf{\Omega})\\ &= \sum_{i=1}^{m}\frac{1}{l_i}\sum_{j=1}^{l_i}\ln(1+\exp(-y^i_j \mathbf{w}_i^T \mathbf{x}^i_j))\\ &+\frac{\lambda}{2}Tr(\mathbf{W}\mathbf{W}^\top) + \frac{\beta}{2}Tr(\mathbf{W}\mathbf{\Omega}^{-1}\mathbf{W}^\top)\\ &+\frac{\gamma}{2}\sum_{i=1}^{m}\frac{1}{n_i}Tr(\mathbf{w}_i^T X_i L_i X_i^T \mathbf{w}_i)\\ &s.t., \quad \mathbf{\Omega} \succeq 0, \quad Tr(\mathbf{\Omega}) = 1.\end{aligned}$$

Under this new semi-supervised unified framework, our proposed SMTL-LLR model can leverage the large amount of unlabeled review data in addition to the labeled ones to learn multiple review spam detection models simultaneously, and then, what is learned for one task can help other related tasks be learned better. In contrast, previous single-task learning based review spam detection models, which are trained independently, and are typically built on a limited set of labeled review data, cannot benefit from this.

3.3 Stochastic Alternating Method

There are two parameters $\mathbf{W}$ and Ω in the objective function of the proposed SMTL-LLR model. It is not easy to optimize the objective function against the two parameters at the same time. We then develop a stochastic alternating method to cope with the optimization problem for SMTL-LLR, i.e., alternatively updating one parameter by fixing the other. In particular, we initialize $\mathbf{W}$ with the values randomly chosen from $[0, 1]$, and initialize $\mathbf{\Omega}$ as a diagonal matrix, where $\mathbf{\Omega}_{ii} = \frac{1}{m}$. For each iteration, the key update steps for the two parameters are described as follows

- *Step 1*: Update $\mathbf{W}$ while $\mathbf{\Omega}$ is fixed.

$$\mathbf{W} \leftarrow \operatorname*{argmin}_{\mathbf{W}} P_{SMTL}(\mathbf{W}, \mathbf{\Omega})$$

- *Step 2*: Update $\mathbf{\Omega}$ while $\mathbf{W}$ is fixed.

$$\Omega \leftarrow \operatorname*{argmin}_{\Omega} P_{SMTL}(\mathbf{W}, \mathbf{\Omega})$$

3.3.1 Updating W While Fixing Ω

For *Step 1* of the alternating optimization method, we introduce a stochastic gradient descent method to efficiently update the parameter $\mathbf{W}$, while Ω is fixed. Formally, given a learning task T_i, we randomly choose a subset or mini-batch of reviews $A_b^i = \{(\mathbf{x}_j^i, y_j^i)|j \in [l_i]\}$ from the labeled set $\mathcal{L}_i$ in a particular iteration, where $[l_i]$ denotes $\{1, \ldots, l_i\}$ and $|A_b^i| = r \ll l_i$. Based on the subset of labeled reviews A_b^i, we can construct an unbiased estimate of the objective function

$$\begin{aligned}
&P_{SMTL}(\mathbf{W}, \mathbf{\Omega}, \{A_b^i\}_{i=1}^m) \\
&= \sum_{i=1}^{m} \frac{1}{r} \sum_{j \in A_b^i} \ln(1 + \exp(-y_j^i \mathbf{w}_i^T \mathbf{x}_j^i)) \\
&+ \frac{\lambda}{2} Tr(\mathbf{W}\mathbf{W}^{\mathbf{T}}) + \frac{\beta}{2} Tr(\mathbf{W}\mathbf{\Omega}^{-1}\mathbf{W}^T) \\
&+ \frac{\gamma}{2} \sum_{i=1}^{m} \frac{1}{n_i} Tr(\mathbf{w}_i^T X_i L_i X_i^T \mathbf{w}_i)
\end{aligned}$$

We can then obtain an unbiased stochastic gradient of the objective

$$\begin{aligned}
&\nabla_{\mathbf{W}} P_{SMTL}(\mathbf{W}, \mathbf{\Omega}, \{A_b^i\}_{i=1}^m) \\
&= [\mathbf{g}_b^1, \ldots, \mathbf{g}_b^m] + \lambda \mathbf{W} + \beta \mathbf{W}\Omega^{-1} \\
&+ [\gamma \frac{1}{n_1} X_1 L_1 X_1^T \mathbf{w}_1, \ldots, \gamma \frac{1}{n_m} X_m L_m X_m^T \mathbf{w}_m],
\end{aligned}$$

where

$$\mathbf{g}_b^i = \frac{1}{r} \sum_{j \in A_b^i} \frac{-y_j^i \mathbf{x}_j^i}{1 + \exp(y_j^i \mathbf{w}_i^T \mathbf{x}_j^i)}.$$

Next, the model parameter $\mathbf{W}$ can be updated via stochastic gradient descent method

$$\mathbf{W}_{t+\frac{1}{2}} = \mathbf{W}_t - \eta_t \nabla_{\mathbf{W}} P_{SMTL}(\mathbf{W}, \mathbf{\Omega}, \{A_b^i\}_{i=1}^m)$$

where $\eta_t > 0$ refers to learning rate in iteration *t*.

Note that, after each update step for the parameter $\mathbf{W}$, we perform a scaling process by forcing the solution

$$\|\mathbf{W}_{t+1}\|_F \leq \sqrt{2m \ln(2)/\lambda},$$

and then have the following update rule

$$\mathbf{W}_{t+1} = \min(1, \frac{\sqrt{2m \ln(2)/\lambda}}{\|\mathbf{W}_{t+\frac{1}{2}}\|_F}) \mathbf{W}_{t+\frac{1}{2}}.$$

We provide a straightforward theoretical analysis, which shows an upper bound of the norm of the optima solution $\mathbf{W}_*$, and explains why we perform the above scaling step. Using the fact that

$$P_{SMTL}(\mathbf{W}_*) \leq P_{SMTL}(0),$$

we thus have

$$\begin{aligned}
\frac{\lambda}{2} \|\mathbf{W}_*\|_F^2 &\leq P_{SMTL}(\mathbf{W}_*) \\
&\leq P_{SMTL}(0) = m \ln(2).
\end{aligned}$$

The fist inequality is guaranteed by

$$\ln(1 + \exp(-y_j^i \mathbf{w}_i^\top \mathbf{x}_j^i)) > 0,$$

$$Tr(\mathbf{W}\mathbf{\Omega}^{-1}\mathbf{W}^\top) \geq 0,$$

and

$$\mathrm{Tr}(\mathbf{w}_i^\top X_i L_i X_i^\top \mathbf{w}_i) \geq 0.$$

3.3.2 Updating Ω While Fixing W

The second step of the stochastic alternating method is equivalent to solving the following optimization problem

$$\begin{aligned}
&\min_{\Omega} \ Tr(W\Omega^{-1}W^\top) \\
&s.t., \ \mathbf{\Omega} \succeq 0, \ Tr(\mathbf{\Omega}) = 1.
\end{aligned}$$

This convex formulation enjoys the following closed-form solution (Zhang and Yeung, 2010)

$$\mathbf{\Omega} = \frac{(\mathbf{W}^\top \mathbf{W})^{\frac{1}{2}}}{Tr((\mathbf{W}^\top \mathbf{W})^{\frac{1}{2}})}.$$

It is obviously observed that Ω models the correlations between each pair of the tasks or the models.

Algorithm 1 summarizes the stochastic alternating optimization method for SMTL-LLR. Given labeled and unlabeled review data for multiple review domains, we run the algorithm for P alternating loops. Within each loop p, we update the model parameter $\mathbf{W}$ for T iterations via stochastic gradient descent method, where *B* is number of mini-batches; after that, we update the task covariance matrix $\mathbf{\Omega}$ once based on new $\mathbf{W}$. The procedure is performed iteratively until it is converged. Then, multiple optimized review spam detection models and task covariance matrix would be learned finally.

Algorithm 1 Stochastic Alternating Method

Input:
Labeled and unlabeled review data for multiple tasks
Initial learning rate η_0, hyper-parameter δ
Regularization parameters λ, β, γ
Initialization:
Initialize $\mathbf{W}$ with values randomly chosen from [0, 1]
Initialize $\mathbf{\Omega} = diag(1/m, \ldots, 1/m)$
for $p = 1, \ldots, P$ **do**
  $\widetilde{\mathbf{W}}_1 = \mathbf{W}$
  for $t = 1, \ldots, T$ **do**
    Learning rate $\eta_t = \frac{\eta_0}{1+\eta_0 \delta t}$
    Randomly shuffle reviews in the training set
    for $b = 1, \ldots, B$ **do**
      Compute $\nabla_{\mathbf{W}} P_{SMTL}(\mathbf{W}, \mathbf{\Omega}, \{A_b^i\}_{i=1}^m)$
      Update $\widetilde{\mathbf{W}}_{t+\frac{1}{2}} = \widetilde{\mathbf{W}}_t$
        $-\eta_t \nabla_{\mathbf{W}} P_{SMTL}(\mathbf{W}, \mathbf{\Omega}, \{A_b^i\}_{i=1}^m)$
      $\widetilde{\mathbf{W}}_{t+1} = \min(1, \frac{\sqrt{2m\ln(2)/\lambda}}{\|\widetilde{\mathbf{W}}_{t+\frac{1}{2}}\|_F}) \widetilde{\mathbf{W}}_{t+\frac{1}{2}}$
    end for
  end for
  Update $\mathbf{W} = \widetilde{\mathbf{W}}_{T+1}$
  Update $\mathbf{\Omega} = \frac{(\mathbf{W}^\top \mathbf{W})^{\frac{1}{2}}}{Tr((\mathbf{W}^\top \mathbf{W})^{\frac{1}{2}})}$
end for
Output: $\mathbf{W}$ and $\mathbf{\Omega}$

In addition, we also rely on the stochastic alternating method to optimize the proposed MTL-LR method. Differently, we need to remove all the terms related to unlabeled data, i.e., discarding the Laplacian regularization term from the objective function and gradient.

4 Experiments

In this section, we evaluate the proposed multi-task learning methods MTL-LR and SMTL-LLR for review spam detection, and demonstrate the improved effectiveness of the methods over other well-established baselines.

4.1 Data Sets

Due to big challenge in manually recognizing deceptive reviews, there are limited benchmark opinion spam data in this field. We used three ground truth data sets from the review domains, *doctor*[2], *hotel*[3], and *restaurant*[4], respectively, to evaluate the proposed methods, which were created by following the similar rules used in (Ott et al., 2011). Then, for each ground truth review data set, we randomly collected a large number of unlabeled reviews (10,000), which were written about the same entities or domain. Table 2 shows some data statistics, where the last column computes the ratio of labeled reviews to unlabeled ones.

	Spam/Nonspam	Unlabeled	Ratio
Doctor	200/200	10,000	4.0%
Hotel	300/300	10,000	6.0%
Restaurant	200/200	10,000	4.0%

Table 2: Some statistics of review data sets.

4.2 Experimental Setup

We followed previous work (Mihalcea and Strapparava, 2009; Ott et al., 2011), and leveraged text unigram and bigram term-frequency features to train our models for review spam detection. This problem setting is quite useful, for example, when user behavior data are sparse or even not available in practical applications.

Supervised classification models, such as logistic regression (LR) and support vector machines (SVM), have been used to identify fake review spam (Jindal and Liu, 2008; Ott et al., 2011). We compared our methods with the two models. Semi-supervised positive-unlabeled (PU) learning was employed for review spam detection, then we chose one representative PU learning method (Liu et al., 2002) to evaluate our models. We did not compare our methods with the two-view co-training method, which was used for fake review detection (Li et al., 2011), because the *reviewer view* data are not available in the ground truth review sets. Instead, we selected a well-known semi-supervised transductive SVM (TSVM) (Joachims, 1999) to evaluate our models. Different from the proposed methods, we trained each of above baselines in a single domain, because they are single-task learning methods. Moreover, we also compared our methods with one well-established multi-task learning baseline MTRL (Zhang and Yeung, 2010), which has not been used

[2]https://www.ratemds.com

[3]https://www.tripadvisor.com

[4]http://www.yelp.com

for review spam detection problem.

It is important to specify appropriate values for the parameters in the proposed methods. In our setting, we used the learning rates η_t that asymptotically decrease with iteration numbers (Bottou, 2012). Following previous work (Ott et al., 2011; Chen and Chen, 2015), we conducted five-fold cross-validation experiments, and determined the values of the regularization and hyper parameters via a grid-search method.

4.3 Experimental Results

Table 3 reports the spam and nonspam review detection accuracy of our methods SMTL-LLR and MTL-LR against all other baseline methods. In terms of 5% significance level, the differences between SMTL-LLR and the baseline methods are considered to be statistically significant.

	Doctor	Hotel	Restaurant	Average
SMTL-LLR	85.4%	88.7%	87.5%	87.2%
MTL-LR	83.1%	86.7%	85.7%	85.2%
MTRL	82.0%	85.4%	84.7%	84.0%
TSVM	80.6%	84.2%	83.8%	82.9%
LR	79.8%	83.5%	83.1%	82.1%
SVM	79.0%	83.5%	82.9%	81.8%
PU	68.5%	75.4%	74.0%	72.6%

Table 3: Spam and nonspam review detection results in the doctor, hotel, and restaurant review domains.

Under symmetric multi-task learning setting, our methods SMTL-LLR and MTL-LR outperform all other baselines for identifying spam reviews from nonspam ones. MTL-LR achieves the average accuracy of 85.2% across the three domains, which is 3.1% and 3.4% better than LR and SVM trained in the single task learning setting, and 1.2% higher than MTRL. Training with a large quantity of unlabeled review data in addition to labeled ones, SMTL-LLR improves the performance of MTL-LR, and achieves the best average accuracy of 87.2% across the domains, which is 3.2% better than that of MTRL, and is 4.3% better than TSVM, a semi-supervised single task learning model. PU gives the worst performance, because learning only with partially labeled positive review data (spam) and unlabeled data may not generalize as well as other methods.

4.4 Performance versus Unlabeled Data Size

Figure 1 plots SMTL-LLR accuracy versus unlabeled data sizes from *0* to *10,000*, where *0* corresponds to using only labeled data to build the model, i.e., MTL-LR. Note that we first randomly sampled *2,000* unlabeled reviews to build the first set, and then created the second set by appending another randomly selected set of *2,000* reviews to the previous one. We repeated the process until all the unlabeled review data sets were created.

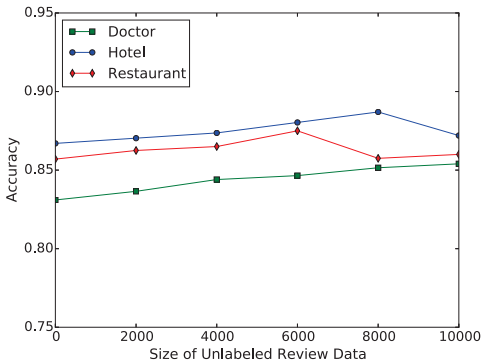

Figure 1: Accuracy versus Unlabeled Data Size.

We observed that learning from unlabeled reviews does help to boost the performance of MTL-LR, which was trained with labeled data alone. The performance of SMTL-LLR improves when training with more and more unlabeled review data. This is because the useful patterns learned from unlabeled data perhaps supports SMTL-LLR to generalize better. But continuing to learn from much more unlabeled reviews may even harm the performance. One explanation is that appending more unlabeled data may also incur noisy information to learning process. Interestingly, the performance of SMTL-LLR keeps increasing on the doctor domain, when training with more and more unlabeled reviews up to *10,000*. From above observations, we conclude that an elaborately selected set of high-quality unlabeled review data may help SMTL-LLR to learn better.

4.5 Task Correlation

Based on the covariance matrix (Ω) learned from the review spam detection tasks, we obtained the correlation between each pair of tasks for doctor, hotel, and restaurant domains, as shown in Table 4. The review spam detection tasks are highly correlated with each other for hotel and restaurant domains (0.772).

This is reasonable due to the large amount of commonality shared between the two domains. We can see that the tasks are also positively correlated between hotel and doctor, as well as between doctor and restaurant domains.

	Doctor	Hotel	Restaurant
Doctor	1.0	0.688	0.638
Hotel	0.688	1.0	0.772
Restaurant	0.638	0.772	1.0

Table 4: Task correlations.

4.6 Shared Text Features among Tasks

Table 5 lists top weighted shared text features among the review spam detection tasks for doctor, hotel, and restaurant domains. Generally, review spammers demonstrate similar motivations when creating deceptive review spam, i.e., promoting their own products/services or defaming those of their competitors. Though different aspects or entities can be commented on across different domains, we find that many features or expressions are indeed shared among the three review domains. As we know, deceptive reviewers normally write up reviews for making money, thus they prefer choosing exaggerated language in their lies, no matter which domains they are working with. As shown in the first row for spam category, they tend to exaggerate their sentiments using the words like "definitely", "sure", "highly", and so on.

In contrast, truthful reviewers contribute reviews for sharing their true feelings or personal anecdotes. They are willing to write up detailed factual experiences, for example, about the doctors they visited or delicious foods they enjoyed. Their reviews thus tend to contain language patterns in past tense, such as "went", "did", and "took" shown in the second row.

5 Conclusions

We have coped with the problem of detecting deceptive review spam. Given the limited labeled review data for individual domains, we formulated it as a multi-task learning problem. We first developed a multi-task learning method via logistic regression (MTL-LR), which allows to boost the learning for one task by sharing the knowledge contained in the training signals of other related tasks. To leverage the unlabeled data, we introduced a graph Laplacian regularizer into each base model, and proposed a semi-supervised multi-task learning model via Laplacian regularized logistic regression (SMTL-LLR). Moreover, to deal with the optimization problem, we developed a stochastic alternating method. Experimental results on real-world review data demonstrated the superiority of SMTL-LLR over several well-established baseline methods.

Labels	Features
Spam	staff, friendly,comfortable, really, right, experience, best, way, amazing, check, away, staff friendly, definitely, sure, highly recommend
Nonspam	good, just, like, went, did, people, excellent, took, wonderful, things, day, fantastic, know, going, nice

Table 5: Top weighted shared text features for spam/nonspam category across the three review domains.

For future work, we plan to create much more ground truth review data from other review domains and different applications like forums or microblogs, and further test our proposed models for deceptive opinion spam detection. We also plan to incorporate our model into a practical opinion mining system, in this way, more reliable opinion and sentiment analysis results can be then expected.

References

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. 2006. Multi-task feature learning. In *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*, pages 41–48, Vancouver, British Columbia, Canada.

Bart Bakker and Tom Heskes. 2003. Task clustering and gating for bayesian multitask learning. *The Journal of Machine Learning Research*, 4:83–99.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100.

Léon Bottou. 1997. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*, pages 421–436.

Rich Caruana. 1997. Multitask learning. In *Machine Learning*, pages 41–75.

Yu-Ren Chen and Hsin-Hsi Chen. 2015. Opinion spam detection in web forum: A real case study. In *Proceedings of the 24th International Conference on World Wide Web*, pages 173–183, Republic and Canton of Geneva, Switzerland.

Francesco De Comite, Francois Denis, Remi Gilleron, and Fabien Letouzey. 1999. Positive and Unlabeled Examples Help Learning. In *Proceedings of the Tenth International Conference on Algorithmic Learning Theory*, Lecture Notes in Artificial Intelligence, pages 219–230, Tokyo, Japan. Springer Verlag.

Theodoros Evgeniou and Massimiliano Pontil. 2004. Regularized multi–task learning. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117, New York, NY, USA.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, pages 171–175.

D. Hernandez, R. Guzman, M. Montes-y-Gomez, and P. Rosso 2013. Using PU-learning to detect deceptive opinion spam. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 38–45, Atlanta, Georgia, USA.

Steven C. H. Hoi, Rong Jin, and Michael R. Lyu. 2006. Large-scale text categorization by batch mode active learning. In *Proceedings of the 15th International Conference on World Wide Web*, pages 633–642, New York, NY, USA.

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the International Conference on Web Search and Data Mining*, pages 219–230, Palo Alto, California, USA.

Thorsten Joachims. 1999. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 200–209, San Francisco, CA, USA.

Fangtao Li, Minlie Huang, Yi Yang, and Xiaoyan Zhu. 2011. Learning to identify review spam. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, page 2488.

Jiwei Li, Myle Ott, Claire Cardie, and Eduard H. Hovy. 2014. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1566–1576, Baltimore, MD, USA.

Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. 2010. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 939–948.

Bing Liu, Wee Sun Lee, Philip S Yu, and Xiaoli Li. 2002. Partially supervised classification of text documents. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 387–394.

Jun Liu, Shuiwang Ji, and Jieping Ye. 2009. Multi-task feature learning via efficient l 2, 1-norm minimization. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 339–348. AUAI Press.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312, Stroudsburg, PA, USA.

Thomas P. Minka. 2003. A comparison of numerical optimizers for logistic regression. Technical report, CMU Technical Report.

Arjun Mukherjee, Bing Liu, and Natalie Glance. 2012. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st International Conference on World Wide Web*, pages 191–200.

Arjun Mukherjee, Abhinav Kumar, Bing Liu, Junhui Wang, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013a. Spotting opinion spammers using behavioral footprints. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 632–640.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 309–319, Portland, Oregon.

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

S. Thrun and J. O'Sullivan. 1996. Discovering structure in multiple learning tasks: The TC algorithm. In L. Saitta, editor, *Proceedings of the 13th International Conference on Machine Learning*, San Mateo, CA. Morgen Kaufmann.

Guan Wang, Sihong Xie, Bing Liu, and Philip S. Yu. 2012. Identify online store review spammers via social review graph. *ACM Trans. Intell. Syst. Technol.*, 3(4):61:1–61:21, September.

Yu Zhang and Dit-Yan Yeung. 2010. A convex formulation for learning task relationships in multi-task learning. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 733–442, Catalina Island, CA, USA.