A Personalized Sentiment Analysis Framework Based on the Exploration of Homogeneity Theory and Technology

AUTHORS
Mohsen Mohammadihadadan, hannaoleszkiewicz
Department of Technical and Engineering, South Tehran Branch, Islamic Azad University, Tehran, IRAN

Abstract:
People have different views on the same subject and thus challenge the traditional sentiment analysis system. Traditional sentiment analysis systems rely on traditional statistical models, which suffer from data scarcity. This paper builds a personalized sentiment analysis system that uses social relationship information to balance personalization and over-provisioning. The method is based on a network lasso that automatically brings together socially nearby users so that users in the same group share the same personalized model. At the same time, a distributed optimization algorithm is designed to make the personalized sentiment analysis system extend to large networks. Experiments on the Yelp review dataset show that our approach is always better than the competition approach.

Keywords: Emotional analysis system, social relationship information, personalized model, network lasso, optimization algorithm

1 Introduction
Sentiment analysis is a fundamental natural language processing (NLP) problem that underlies numerous tasks such as marketing analysis and personalized recommendation. One challenge in sentiment analysis is that people often express different opinions toward the same topic, depending on their language habit, personal character, and opinion bias. Traditional approaches usually employ population-level classification: predicting sentiment polarities based on models trained with labeled instances of all users. The setup results in inaccurate analysis for many individuals, especially for long-tail users with opinions differ from the general tendencies. We can overcome this problem by incorporating individual-level information. However, naively training different classifiers for different authors usually yields unsatisfactory results, due to the limited availability of user-specific opinionated data. Previous work addresses this challenge by employing personalization techniques adapted from information retrieval or speech recognition (Leggetter and Woodland, 1995). For users with insufficient personalization data, the basic idea is that we can utilize a model estimated on all the labeled data or the data from other similar users (Al Boni et al., 2015; Song et al., 2015). One way to group similar users together is to exploit social network information and cluster individuals based on social distances between users.

The theory of homophily (McPherson et al., 2001) suggests that socially connected individuals tend to have similar behaviors or share similar interests. Applying the theory in the context of sentiment analysis, we assume socially linked users are more likely to express similar sentiment meanings toward the same words and topics. The assumption helps to alleviate the data scarcity of individual-level sentiment analysis.
scarcity issue and smooth personalized models over social connections. Previous studies on incorporating social relations into sentiment analysis have attained some successes on microblog data (Song et al., 2015). However, the approaches are often computationally expensive (Song et al., 2015) or generalize poorly to long-tail users (Vosoughi et al., 2016). In this work, we present a flexible and powerful framework to leverage social relation information for sentiment analysis. The system is based on network lasso (Hallac et al., 2015), a generalization of the group lasso to a network setting that allows for simultaneous clustering and optimization on graphs. To deal with large networks with millions of nodes, we develop an algorithm based on the Alternating Direction Method of Multipliers (ADMM) (Boyd et al., 2011) to solve this optimization problem in a distributed and scalable manner. Instead of backing off personalized models estimated with insufficient data to a global model, network lasso can automatically group long-tail users to a larger cluster and estimate models accordingly. This mechanism reduce the variance of the personalized models, while the models can still benefit from personalized information. Experiments on Yelp review datasets show that our approach consistently outperform competitive methods.

2 Related Work

Early work on sentiment analysis focused on review data, and several supervised machine learning based classification approaches were proposed to automatically categorize a user review to positive or negative class. Pang et al. (2002) employed different machine learning techniques such as Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machine (SVM) for binary sentiment classification of movie reviews. McDonald et al. (2007) presented a structured model for jointly classifying the sentiment at sentence and document level. Recent work has been focused on sentiment analysis on short and noisy text data, especially on classifying the Twitter messages into positive, negative, and neutral categories (Go et al., 2009; Kouloumpiset al., 2011). Ortigosa et al. (2014) perform sentiment classification and sentiment change detection on Facebook comments using lexicon and machine learning based approaches. Boyd et al. (2011) propose a target-dependent and contextual based approach to perform sentiment classification of tweets. As the texts are short, learning expressive representations using neural networks have been shown to be very effective. Vosoughi et al. (2016) present a model that learns tweet embedding using character-level CNN-LSTM encoder-decoder, which are trained with three million, randomly selected English tweets. A hierarchical LSTM model is proposed to model rich contexts in tweet, particularly some long-range contexts (Vosoughi et al., 2016).

Most existing work follows the standard classification setup, and estimates a single statistical model for all the users. As different people may hold different opinions toward the same topic, ignoring personalized information may lead to suboptimal solutions for sentiment classification. Some efforts have been made to adopt personalization techniques such as linear transformation of model parameters (Leggetter and Woodland, 1995) and collaborative filtering (Breese et al., 1998) to sentiment analysis (Al Boni et al., 2015; Song et al., 2015). In order to address the data scarcity problem, there are a few attempts to incorporate social link information into sentiment classification. Kouloumpiset al. (2011) model social relations using
the graph Laplacian of the adjacency graph representation of the social network, and socially-
similar users are encouraged to have similar labels. They leverage a similar intuition, but using a
factor graph based approach.

Our method is closely related to that proposed by Vosoughi et al. (2016), which define
personalized models in addition to a population-level model. For users with a small number of
labeled instances (i.e., long-tail users), this approach will backoff from the personalized models
to the population-level model. We do not assume that there is an underlying model space
shared by everyone, so that our method can benefit personalized information for long-tail users
better. As we will show late, this design choice results in significant improvements for
personalized sentiment classification.

3 Model
In this section, we first describe the general framework to incorporate social network structures
for personalized sentiment analysis. The framework can work with different types of social
networks as well as different loss functions. Then, we will present a distributed optimization
algorithm based on the Alternating Direction Method of Multipliers (ADMM; Boyd et al., 2011).
The algorithm is guaranteed to converge to a global optimum even on very large networks

3.1 Network Lasso for Personalized Sentiment Analysis

Given a user \( u \) and a set of labeled messages \( \{ y_{ij} \} \) that are written by the user, the
personalized sentiment classifier for \( u \) can be obtained by solving the following optimization
problem:

\[
\minimize_{(x_i, y_i) \in D_u} \ell(y_i, f(x_i, w_i))
\]  

(1)

where \( x_i \) is the message representation that can be a simple bag-of-word vector, \( x_i \) is the
message-level sentiment label, and \( w_i \) is the personalized model weights for the user. We can
adopt different types of loss functions, and we consider hinge loss

(i.e., \( \ell(y_i, f(x_i, w_i)) = \left[ 1 - y_i w_i^T x_i \right]^+ \)) and log loss

(i.e., \( \ell(y_i, f(x_i, w_i)) = \log(1 + \exp(1 - y_i w_i^T x_i)) \)) in this work.

The personalized models that are independent with each other usually lead to unsatisfactory
results, due to the insufficient training data for each individual user. We can overcome this
problem by incorporating social network information. In particular, we assume that if user \( u \) is
socially connected with user \( u' \), they should express similar sentiment to the same topic, and
therefore \( w_u \) should be close to \( w_{u'} \). To ensure this, we can simply add a term

\( \| w_u - w_{u'} \|_2 \) to the loss function. Formally, we seek to optimize a set of variables \( \{ w_u, \ldots, w_{u'} \} \)
for the \( u \) users, which corresponds to the following problem,

\[
\minimize_{i=1}^{n} f_i(w_i) + \lambda \sum_{(j,k) \in E} \| w_j - w_k \|_2
\]  

(2)
\[ f_i(w_i) = \sum_{(x_i, y_i) \in D_i} \ell(y_i, f(x_i, w_i)) \]  

(3)

where \( D_i \) are the set of social links between the users, and \( f_i(w_i) \) is the loss function that captures the personalized information of user \( i \).

The \( \ell_1 \)-norm penalty over the personalized model difference,\( \gamma - \gamma_i \), defines the network lasso problem. Similar to the group lasso problem, it encourages the differences between connected nodes to be exactly zero, rather than just close to zero. This leads to several groups of users, where each group has exact the same solution for the variable \( \gamma_i \). Increasing \( \lambda \) yields to larger sizes of these clusters until all the users are grouped into the same cluster, and tuning \( \lambda \) offers the flexibility to better fit different datasets.

3.2 ADMM Optimization

We can solve the above network lasso problem using standard interior point methods on datasets with small social networks. However, there are often hundreds of millions of users in real world social media platforms such as Twitter, Facebook, and Yelp. In order to make our sentiment analysis system scalable to these large social networks, a distributed optimizer solution is necessary so that computational and storage limits do not constrain the adoption of the system to new datasets and settings. We choose to utilize the Alternating Direction Method of Multipliers (ADMM; Boyd et al., Nguyen et al.), a well-known method for distributed convex optimization, to find the solution for our personalized sentiment classification problem.

We can partition the data instances into \( n \) groups, and each group of data will be processed by a single machine in the distributed setting. Specifically, each individual machine solves its own private objective function that involved certain users, passes this solution to other machines that involved neighbors of these users, and repeats the process until the overall process converges. To do this, we need to introduce a copy of \( \gamma_i \) at each edge \((i, j)\), which we call \( \gamma_{ij} \).

Let \( N(i) \) be the set of nodes connecting to \( \gamma_i \), the original network lasso problem can be rewritten as an equivalent problem,

Minimize \[ \sum_{i=1}^{n} f_i(w_i) + \lambda \sum_{(j, k) \in E} \| w_j - w_k \|_2 \]

subject to \( \gamma_i = \gamma, i = 1, \ldots, m, j \in N(i) \).

(4)

We can derive the augmented Lagrangian form for the problem,

\[
L(w, v, u) = \sum_{i=1}^{n} f_i(w_i) + \sum_{(j, k) \in E} (\lambda \| v_{jk} - v_{kj} \|_2 \\
+ \frac{\rho}{2} (\| w_j - v_{jk} + u_{jk} \|_2^2 + \| w_k - v_{jk} + u_{kj} \|_2^2)) - \frac{\rho}{2} (\| u_{jk} \|_2^2 + \| u_{kj} \|_2^2),
\]

(5)

Where \( u \) scaled dual variable and \( \rho > 0 \) is the penalty parameter. ADMM iteratively optimize these parameters, and we will use superscript \( t \) to denote the iteration number. The update
steps for each iteration are:

\[ w^{t+1} = \arg \min_w L(w, v^t, u^t) \]

\[ v^{t+1} = \arg \min_v L(w^{t+1}, v, u^t) \]

\[ u^{t+1} = u^t + (w^{t+1} - v^{t+1}) \]  \hspace{1cm} (6)

To update \( w_i \) we can minimize a separable sum of functions for each user. Thus, it can be computed independently and solved in parallel. For user \( i \), the update rule is

\[ w_i^{t+1} = \arg \min_{w_i} (f_i(w_i) + \frac{\rho}{2} \sum_{j \in N(i)} \|w_i - v^{t+1}_{jk} + u^{t+1}_{jk}\|_2^2) \]  \hspace{1cm} (7)

We can also update \( v \) separately with respect to each social connection, which makes distributed optimization trivial. Note that we need to jointly update \( \lambda \) and \( \theta \) for edge \((i, j)\)

\[ v^{t+1}_{ij}, v^{t+1}_{ji} = \arg \min_{v_{ij}, v_{ji}} (\lambda \|v_{ij} - v_{ji}\|_2^2 + \frac{\rho}{2} (\|w_i - v^{t+1}_{ij} + u^{t+1}_{ij}\|_2^2 + \|w_j - v^{t+1}_{ji} + u^{t+1}_{ji}\|_2^2)) \]  \hspace{1cm} (8)

As shown by Hallac et al. (2015), the problem has a closed-form analytical solution, which has the form

\[ v_{ij} = \theta (w_i + u_j) + (1 - \theta) (w_j + u_j) \]

\[ v_{ji} = (1 - \theta) (w_i + u_j) + \theta (w_j + u_j) \]

where

\[ \theta = \max(1 - \lambda \rho (\|w_i + u_j\| - (w_j + u_j)\|_2^2))^{-1} \]  \hspace{1cm} (9)

Finally, the \( u \)-Update step is straightforward, which can be calculated in parallel as well,

\[ u^{t+1}_{ji} = u^{t}_{ji} + (w^{t+1}_{ij} - v^{t+1}_{ji}) \]  \hspace{1cm} (10)

4 Experiments

In this section, we present our experimental findings. Our experiments are conducted using the Yelp review dataset from the Yelp Dataset Challenge. We will first briefly describe our preprocessing steps, and then present our main experiment and analysis results.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Author</th>
<th># Relation</th>
<th># Positive</th>
<th># Negative</th>
<th># Review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yelp 2014</td>
<td>182,442</td>
<td>382,383</td>
<td>330,737</td>
<td>107,345</td>
<td>438,082</td>
</tr>
<tr>
<td>Yelp 2015</td>
<td>261,601</td>
<td>436,468</td>
<td>452,481</td>
<td>150,044</td>
<td>602,525</td>
</tr>
<tr>
<td>Yelp 2016</td>
<td>201,103</td>
<td>241,608</td>
<td>302,573</td>
<td>98,612</td>
<td>401,185</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the Yelp product review datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Yelp 2014</th>
<th>Yelp 2015</th>
<th>Yelp 2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>92.47</td>
<td>93.07</td>
<td>93.18</td>
</tr>
<tr>
<td>Support Machine</td>
<td>Vector</td>
<td>91.51</td>
<td>92.21</td>
</tr>
</tbody>
</table>

IJMRAS-ISSN2640-7272, Mohsen Mohammadihadadan, hannaoleszkiewicz
Table 2: Average F1 score on the Yelp test sets. The best results are in bold.

4.1 Data
We evaluate our models using Yelp Academic Data Set released for the Round 9 of the Yelp Dataset Challenge. It is a large-scale dataset consisting of 4.1M reviews and 947K tips by 1M users for 144K businesses. We follow prior work and further split the data into three subsets: Yelp 2014, Yelp 2015, Yelp 2016, according to the post dates of the reviews. The rating information in discrete five-star range is available for the reviews, which can be used as the ground truth label information for the reviews. We treated reviews with more than three stars as positive, and less than three stars as negative.

A Yelp user can be friends of other users, indicating that the user acquaints or trusts his connections. We used the friendship information and created a social network for each of the Yelp 2014, Yelp 2015, Yelp 2016 datasets. The networks are treated as undirected graphs. The statistics of the datasets are presented in Table 1. We randomly split each dataset into training (70%), development (10%) and test (20%) sets.

4.2 Experimental Settings
We train all the models on the training sets, tune parameters on the development sets, and report experiment results on the test sets. As shown in Table 1, the datasets have highly skewed class distributions toward the positive polarity. Thus, we use the average F1 score of positive and negative classes as the evaluation metric.

Competitive systems We compare our proposed methods against standard classification models as well as the personalized sentiment classification models by Vosoughi et al. (2016). The first two systems Logistic Regression and Support Vector Machine (SVM) ignore social relation information, and estimate population-level models with all available data. PMSC-Log and PMSC-Hinge are two systems based on the personalized classification approach proposed by Vosoughi et al. (2016). Here we consider two variants of the approach with log loss and hinge loss functions respectively. Finally, we include NetLasso-Log and NetLasso-Hinge, our proposed sentiment classification systems based on Network Lasso. Both PMSC and NetLass are able to incorporate social relation information to smooth over personalized models and improve overall performance. The systems are trained with bag-of-word representations that are the top 10,000 most frequent words in each dataset.

4.3 Results
One potential issue with network lasso is that it may work poorly for a network with a lot of isolated components, as a node in an isolated component can only communicate with other local nodes. To solve this problem, we expand the network by adding random bridge links between different components until there is a single connected component. Similar smoothing techniques have been widely adopted in machine learning (Rennie et al., 2003) and natural
language processing (Chen and Goodman, 1996) literature.
The evaluation results are presented in Table 2, and the best results are shown in bold font. As shown, our proposed network lasso based personalized sentiment classification systems consistently outperform other baseline systems. Compared with PMSC, network lasso is able to leverage personalized information more effectively by automatically grouping users into clusters. In terms of loss functions, logistic loss works marginally better than hinge loss in all the cases. As the numbers of the training instances are relatively large, the two loss functions tend to yield similar performance (McDonald et al., 2007).

5 Conclusion
This paper presents a personalized sentiment analysis framework that leverages social relation information to overcome the data scarcity problem and improve the overall accuracies of sentiment classification. Our approach is based on network lasso—a technique that explores homophily theory and assumes socially proximate users are more likely to express similar sentiments toward the same topic. We propose to optimize the models in a distributed setting using ADMM, permitting to work with very large networks. We conduct experiments with the Yelp product review dataset, and the results show that our method performs consistently better than competitive baselines. In the future, we would like to adopt our framework to other datasets and social networks, such as Facebook and Twitter.

References
1https://www.yelp.com/dataset_challenge