Model Adaptation for Personalized Opinion Analysis

Mohammad Al Boni[1], Keira Qi Zhou[1], Hongning Wang[2], and Matthew S. Gerber[1]

[1]Department of Systems and Information Engineering
[2]Department of Computer Science
[1,2]University of Virginia, USA
[1,2]{ma2sm,qz4aq,hw5x,msg8u}@virginia.edu

Abstract

Humans are idiosyncratic and variable: towards the same topic, they might hold different opinions or express the same opinion in various ways. It is hence important to model opinions at the level of individual users; however it is impractical to estimate independent sentiment classification models for each user with limited data. In this paper, we adopt a model-based transfer learning solution – using linear transformations over the parameters of a generic model – for personalized opinion analysis. Extensive experimental results on a large collection of Amazon reviews confirm our method significantly outperformed a user-independent generic opinion model as well as several state-of-the-art transfer learning algorithms.

1 Introduction

The proliferation of user-generated opinionated text data has fueled great interest in opinion analysis (Pang and Lee, 2008; Liu, 2012). Understanding opinions expressed by a population of users has value in a wide spectrum of areas, including social network analysis (Bodendorf and Kaiser, 2009), business intelligence (Gamon et al., 2005), marketing analysis (Jansen et al., 2009), personalized recommendation (Yang et al., 2013) and many more.

Most of the existing opinion analysis research focuses on population-level analyses, i.e., predicting opinions based on models estimated from a collection of users. The underlying assumption is that users are homogeneous in the way they express opinions. Nevertheless, different users may use the same words to express distinct opinions. For example, the word "expensive" tends to be associated with negative sentiment in general, although some users may use it to describe their satisfaction with a product's quality. Failure to recognize this difference across users will inevitably lead to inaccurate understanding of opinions.

However, due to the limited availability of user-specific opinionated data, it is impractical to estimate independent models for each user. In this work, we propose a transfer learning based solution, named LinAdapt, to address this challenge. Instead of estimating independent classifiers for each user, we start from a generic model and adapt it toward individual users based on their own opinionated text data. In particular, our key assumption is that the adaptation can be achieved via a set of linear transformations over the generic model's parameters. When we have sufficient observations for a particular user, the transformations will push the adapted model towards the user's personalized model; otherwise, it will back off to the generic model. Empirical evaluations on a large collection of Amazon reviews verify the effectiveness of the proposed solution: it significantly outperformed a user-independent generic model as well as several state-of-the-art transfer learning algorithms.

Our contribution is two-fold: 1) we enable efficient personalization of opinion analysis via a transfer learning approach, and 2) the proposed solution is general and applicable to any linear model for user opinion analysis.

2 Related Work

Sentiment Analysis refers to the process of identifying subjective information in source materials (Pang and Lee, 2008; Liu, 2012). Typical tasks include: 1) classifying textual documents into positive and negative polarity categories, (Dave et al., 2003; Kim and Hovy, 2004); 2) identifying textual topics and their associated opinions (Wang et al., 2010; Jo and Oh, 2011); and 3) opinion summarization (Hu and Liu, 2004; Ku et al., 2006). Approaches for these tasks focus on population-level opinion analyses, in which one model is shared across all users. Little effort has been devoted to personalized opinion analyses, where each user has a particular model, due to the absence of user-

specific opinion data for model estimation.

Transfer Learning aims to help improve predictive models by using knowledge from different but related problems (Pan and Yang, 2010). In the opinion mining community, transfer learning is used primarily for domain adaptation. Blitzer et al. (2006) proposed structural correspondence learning to identify the correspondences among features between different domains via the concept of pivot features. Pan et al. (2010) propose a spectral feature alignment algorithm to align domain-specific sentiment words from different domains for sentiment categorization. By assuming that users tend to express consistent opinions towards the same topic over time, Guerra et al. (2011) applied instance-based transfer learning for real time sentiment analysis.

Our method is inspired by a personalized ranking model adaptation method developed by Wang et al. (2013). To the best of our knowledge, our work is the first to estimate user-level classifiers for opinion analysis. By adapting a generic opinion classification model for each user, heterogeneity among their expressions of opinions can be captured and it help us understand users' opinions at a finer granularity.

3 Linear Transformation Based Model Adaptation

Given a generic sentiment classification model $y = f^s(x)$, we aim at finding an optimal adapted model $y = f^u(x)$ for user u, such that $f^u(x)$ best captures u's opinion in his/her generated textual documents $D^u = \{x_d, y_d\}_{d=1}^{|D|}$, where x_d is the feature vector for document d, y_d is the sentiment class label (e.g., positive v.s., negative). To achieve so, we assume that such adaptation can be performed via a series of linear transformations on $f^s(x)$'s model parameter w_s. This assumption is general and can be applied to a wide variety of sentiment classifiers, e.g., logistic regression and linear support vector machines, as long as they have a linear core function. Therefore, we name our proposed method as LinAdapt. In this paper, we focus on logistic regression (Pang et al., 2002); but the proposed procedures can be easily adopted for many other classifiers (Wang et al., 2013).

Our global model $y = f^s(x)$ can be written as,

$$P^s(y_d = 1|x_d) = \frac{1}{1 + e^{-w^{s\mathsf{T}} x_d}} \quad (1)$$

where w^s are the linear coefficients for the corresponding document features.

Standard linear transformations, i.e., scaling, shifting and rotation, can be encoded via a $V \times (V+1)$ matrix A^u for each user u as:

$$\begin{pmatrix} a^u_{g(1)} & c^u_{g(1),12} & c^u_{g(1),13} & 0 & 0 & b^u_{g(1)} \\ c^u_{g(2),21} & a^u_{g(2)} & c^u_{g(2),23} & \cdots & 0 & b^u_{g(2)} \\ c^u_{g(3),31} & c^u_{g(3),32} & a^u_{g(3)} & \ddots & \vdots & b^u_{g(3)} \\ 0 & \cdots & \cdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & a^u_{g(V)} & b^u_{g(V)} \end{pmatrix}$$

where V is the total number of features.

However, the above transformation introduces $O(V^2)$ free parameters, which are even more than the number of free parameters required to estimate a new logistic regression model. Following the solution proposed by Wang et al. (2013), we further assume the transformations can be performed in a group-wise manner to reduce the size of parameters in adaptation. The intuition behind this assumption is that features that share similar contributions to the classification model are more likely to be adapted in the same way. Another advantage of feature grouping is that the feedback information will be propagated through the features in the same group while adaptation; hence the features that are not observed in the adaptation data can also be updated properly.

We denote $g(\cdot)$ as the feature grouping function, which maps V original features to K groups, and a^u_k, b^u_k and c^u_k as the scaling, shifting and rotation operations over w^s in group k for user u. In addition, rotation is only performed for the features in the same group, and it is assumed to be symmetric, i.e., $c^u_{k,ij} = c^u_{k,ji}$, where $g(i) = k$ and $g(j) = k$. As a result, the personalized classification model $f^u(x)$ after adaptation can be written as,

$$P^u(y_d = 1|x_d) = \frac{1}{1 + e^{-(A^u \tilde{w}^s)^{\mathsf{T}} x_d}} \quad (2)$$

where $\tilde{w}^s = (w^s, 1)$ to accommodate the shifting operation.

The optimal transformation matrix A^u for user u can be estimated by maximum likelihood estimation based on user u's own opinionated document collection D^u. To avoid overfitting, we penalize the transformation which increases the discrepancy between the adapted model and global model by the following regularization term,

$$R(A^u) = -\frac{\eta}{2}\sum_{k=1}^{K}(a^u_k - 1)^2 - \frac{\sigma}{2}\sum_{k=1}^{K} b^{u\,2}_k - \frac{\epsilon}{2}\sum_{k=1}^{K}\sum_{i,g(i)=k}\sum_{j\neq i,g(j)=k} {c^u_{k,ij}}^2, \quad (3)$$

where η, σ and ϵ are trade-off parameters controlling the balance among shifting, scaling and rotation operations in adaptation.

Combining the newly introduced regularization term for A^u and log-likelihood function for logistic regression, we get the following optimization problem to estimate the adaptation parameters,

$$\max_{A^u} L(A^u) = L_{LR}(D^u; P^u) + R(A^u) \quad (4)$$

where $L_{LR}(D^u; P^u)$ is the log-likelihood of logistic regression on collection D^u, and P^u is defined in Eq (2).

Gradient-based method is used to optimize Eq (4), in which the gradient for a_k^u, b_k^u and c_k^u can be calculated as,

$$\frac{\partial L(A^u)}{\partial a_k} = \sum_{d=1}^{D^u} \{y_d[1 - p(y_d|x_d)] \sum_{i,g(i)=k} w_i^s x_{di}\} - \eta(a_k - 1)$$

$$\frac{\partial L(A^u)}{\partial b_k} = \sum_{d=1}^{D^u} \{y_d[1 - p(y_d|x_d)] \sum_{i,g(i)=k} x_{di}\} - \sigma b_k$$

$$\frac{\partial L(A^u)}{\partial c_{k,ij}} = \sum_{d=1}^{D^u} \{y_d[1 - p(y_d|x_d)] w_j^s x_{di}\} - \epsilon c_{k,ij}$$

4 Experiments and Discussion

We performed empirical evaluations of the proposed LinAdapt algorithm on a large collection of product review documents. We compared our approach with several state-of-the-art transfer learning algorithms. In the following, we will first introduce the evaluation corpus and baselines, and then discuss our experimental findings.

4.1 Data Collection and Baselines

We used a corpus of Amazon reviews provided on Stanford SNAP website by McAuley and Leskovec. (2013). We performed simple data preprocessing: 1) annotated the reviews with ratings greater than 3 stars (out of total 5 stars) as positive, and others as negative; 2) removed duplicate reviews; 3) removed reviewers who have more than 1,000 reviews or more than 90% positive or negative reviews; 4) chronologically ordered the reviews in each user. We extracted unigrams and bigrams to construct bag-of-words feature representations for the review documents. Standard stopword removal (Lewis et al., 2004) and Porter stemming (Willett, 2006) were applied. Chi-square and information gain (Yang and Pedersen, 1997) were used for feature selection and the union of the resulting selected features are used in the final controlled vocabulary. The resulting evaluation data set contains 32,930 users, 281,813 positive reviews, and 81,522 negative reviews, where each review is represented with 5,000 text features with TF-IDF as the feature value.

Our first baseline is an instance-based adaptation method (Brighton and Mellish, 2002). The k-nearest neighbors of each testing review document are found from the shared training set for personalized model training. As a result, for each testing case, we are estimating an independent classification model. We denote this method as "ReTrain." The second baseline builds on the model-based adaptation method developed by Geng et al. (2012). For each user, it enforces the adapted model to be close to the global model via an additional L2 regularization when training the personalized model. But the full set of parameters in logistic regression need to estimated during adaptation. We denote this method as "Reg-LR."

In our experiments, all model adaptation is performed in an online fashion: we first applied the up-to-date classification model on the given testing document; evaluated the model's performance with ground-truth; and used the feedback to update the model. Because the class distribution of our evaluation data set is highly skewed (77.5% positive), it is important to evaluate the adapted models' performance on both classes. In the following comparisons, we report the average F-1 measure of both positive and negative classes.

4.2 Comparison of Adaptation Performance

First we need to estimate a global model for adaptation. A typical approach is to collect a portion of historical reviews from each user to construct a shared training corpus (Wang et al., 2013). However, this setting is problematic: it already exploits information from every user and does not reflect the reality that some (new) users might not exist when training the global model. In our experiment, we isolated a group of random users for global model training. In addition, since there are multiple categories in this review collection, such as book, movies, electronics, etc, and each user might discuss various categories, it is infeasible to balance the coverage of different categories in global model training by only selecting the users. As a result, we vary the number of reviews in each domain from the selected training users to estimate the global model. We started with 1000 reviews from the top 5 categories (Movies & TV, Books, Music, Home & Kitchen, and Video Games), then evaluated the global model on 10,000 testing users which consist of three groups: light users with 2 to 10 reviews, medium users with 11 to 50 reviews, and heavy users with 51 to 200 reviews. After each evaluation run, we added an extra 1000 reviews and repeated the training and evaluation.

Table 1: Global model training with varying size of training corpus.

Model	Metric	1000	2000	3000	4000	5000
Global	Pos F1	0.741	0.737	0.738	0.734	0.729
	Neg F1	0.106	0.126	0.125	0.132	**0.159**
LinAdapt	Pos F1	0.694	0.693	0.692	0.694	**0.696**
	Neg F1	0.299	0.299	0.296	0.299	**0.304**

Table 2: Effect of feature grouping in LinAdapt.

Method	Metric	100	200	400	800	1000
Rand	Pos F1	0.691	0.692	0.696	0.686	0.681
	Neg F1	0.295	0.298	0.300	0.322	0.322
SVD	Pos F1	0.691	0.698	0.704	0.697	0.696
	Neg F1	0.298	0.302	0.300	0.322	**0.334**
Cross	Pos F1	**0.701**	**0.702**	**0.705**	**0.700**	0.696
	Neg F1	0.298	0.299	**0.303**	**0.328**	0.331

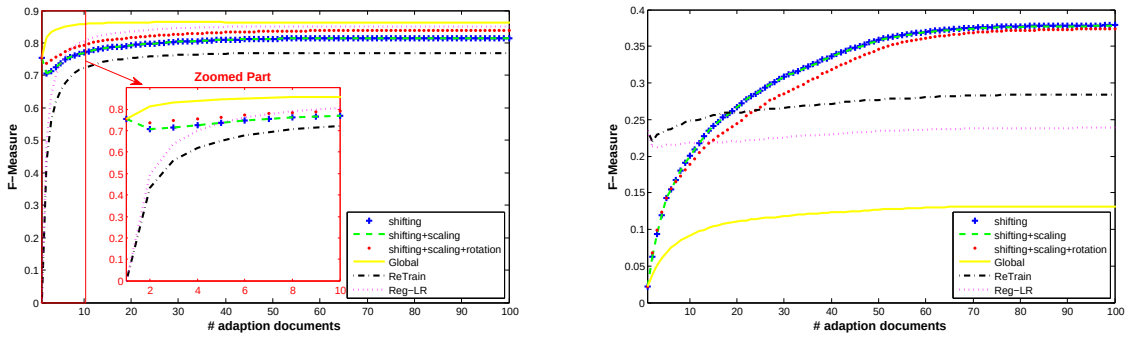
To understand the effect of global model training in model adaptation, we also included the performance of LinAdapt, which only used shifting and scaling operations and $Cross$ feature grouping method with $k = 400$ (detailed feature grouping method will be discussed in the next experiment). Table 1 shows the performance of the global model and LinAdapt with respect to different training corpus size. We found that the global model converged very quickly with around 5,000 reviews, and this gives the best compromise for both positive and negative classes in both global and adaptaed model. Therefore, we will use this global model for later adaptation experiments.

We then investigated the effect of feature grouping in LinAdapt. We employed the feature grouping methods of SVD and $Cross$ developed by Wang et al. (2013). A random feature grouping method is included to validate the necessity of proper feature grouping. We varied the number of feature groups from 100 to 1000, and evaluated the adapted models using the same 10,000 testing users from the previous experiment. As shown in Table 2, $Cross$ provided the best adaptation performance and random is the worse; a moderate group size balances performance between positive and negative classes. For the remaining experiments, we use the $Cross$ grouping with $k = 400$ in LinAdapt. In this group setting, we found that the average number of features per group is 12.47 while the median is 12, which means that features are normally distributed across different groups.

Next, we investigated the effect of different linear operations in LinAdapt, and compared LinAdapt against the baselines. We started LinAdapt with only the shifting operation, and then included scaling and rotation. To validate the necessity of personalizing sentiment classification models, we also included the global model's performance in Figure 1. In particular, to understand the longitudinal effect of personalized model adaptation, we only used the heavy users (4,021 users) in this experiment. The results indicate that the adapted models outperformed the global model in identifying the negative class; while the global model performs the best in recognizing positive reviews. This is due to the heavily biased class distribution in our collection: global model puts great emphasis on the positive reviews; while the adaptation methods give equal weights to both positive and negative reviews. In particular, in LinAdapt, scaling and shifting operations lead to satisfactory adaptation performance for the negative class with only 15 reviews; while rotation is essential for recognizing the positive class.

To better understand the improvement of model adaptation against the global model in different types of users, we decomposed the performance gain of different adaptation methods. For this experiment, we used all the 10,000 testing users: we used the first 50% of the reviews from each user for adaptation and the rest for testing. Table 3 shows the performance gain of different algorithms under light, medium and heavy users. For the heavy and medium users, which only consist 0.1% and 35% of the total population in our data set, our adaptation model achieved the best improvement against the global model compared with Reg-LR and ReTrain. For the light users, who cover 64.9% of the total population, LinAdapt was able to improve the performance against the global model for the negative class, but Reg-LR and ReTrain had attained higher performance. For the positive class, none of those adaptation methods can improve over the global model although they provide a very close performance (in LinAdapt, the differences are not significant). The significant improvement in negative class prediction from model adaptation is encouraging considering the biased distribution of classes, which results in poor performance in the global model.

The above improved classification performance indicates the adapted model captures the heterogeneity in expressing opinions across users. To verify this, we investigated textual features whose sentiment polarities are most/least frequently updated across users. We computed the variance of the absolute difference between the learned feature weights in LinAdapt and global model. High variance indicates the word's sentiment polarity frequently changes across different users. But there are two reasons for a low variance: first, a rare



(a) Positive F-1 measure

(b) Negative F-1 measure

Figure 1: Online adaptation performance comparisons.

Table 3: User-level performance gain over global model from ReTrain, Reg-LR and LinAdapt.

Method	User Class	Pos F1	Neg F1
ReTrain	Heavy	-0.092	0.155*
	Medium	-0.095	0.235*
	Light	-0.157*	**0.255***
Reg-LR	Heavy	-0.010	0.109*
	Medium	-0.005	0.206*
	Light	-0.060	0.232*
LinAdapt	Heavy	-0.046	**0.248***
	Medium	-0.049	**0.235***
	Light	-0.091	0.117*

* p-value< 0.05 with paired t-test.

Table 4: Top 10 words with the highest and lowest variance of learned polarity in LinAdapt.

Variance	Features		
Highest	waste	good	attempt
	money	return	save
	poor	worst	annoy
Lowest	lover	correct	pure
	care	the product	odd
	sex	evil	less than

word that is not used by many users; second, a word is being used frequently, yet, with the same polarity. We are only interested in the second case. Therefore, for each word, we compute its user frequency (UF), i.e., how many unique users used this word in their reviews. Then, we selected 1000 most popular features by UF, and ranked them according to the variance of learned sentiment polarities. Table 4 shows the top ten features with the highest and lowest polarity variance.

We inspected the learned weights in the adapted models in each user from LinAdapt, and found the words like *waste, poor*, and *good* share the same sentiment polarity as in the global model but different magnitudes; while words like *money, instead*, and *return* are almost neutral in global model, but vary across the personalized models. On the other hand, words such as *care, sex, evil, pure*, and *correct* constantly carry the same sentiment across users. Table 5 shows the detailed range of learned polarity for three typical opinion words in 10,000 users. This result indicates LinAdapt well captures the fact that users express opinions differently even with the same words.

Table 5: Learned sentiment polarity range of three typical words in LinAdapt.

Feature	Range	Global Weight	Used as Positive	Used as Negative
Experience	[-0.231,0.232]	0.002	3348	1503
Good	[-0.170,0.816]	0.032	8438	1088
Money	[-0.439,0.074]	-0.013	646	6238

5 Conclusion and Future Work

In this paper, we developed a transfer learning based solution for personalized opinion mining. Linear transformations of scaling, shifting and rotation are exploited to adapt a global sentiment classification model for each user. Empirical evaluations based on a large collection of opinionated review documents confirm that the proposed method effectively models personal opinions. By analyzing the variance of the learned feature weights, we are able to discover words that hold different polarities across users, which indicates our model captures the fact that users express opinions differently even with the same words. In the future, we plan to further explore this linear transformation based adaptation from different perspectives, e.g., sharing adaptation operations across users or review categories.

6 Acknowledgements

This research was funded in part by grant W911NF-10-2-0051 from the United States Army Research Laboratory. Also, Hongning Wang is partially supported by the Yahoo Academic Career Enhancement Award.

References

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128. Association for Computational Linguistics.

Freimut Bodendorf and Carolin Kaiser. 2009. Detecting opinion leaders and trends in online social networks. In *Proceedings of the 2nd ACM workshop on Social web search and mining*, pages 65–68. ACM.

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

Pedro Henrique Calais Guerra, Adriano Veloso, Wagner Meira Jr, and Virgílio Almeida. 2011. From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158. ACM.

Kushal Dave, Steve Lawrence, and David M Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM.

Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger. 2005. Pulse: Mining customer opinions from free text. In *Advances in Intelligent Data Analysis VI*, pages 121–132. Springer.

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

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188.

Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 815–824. ACM.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics.

Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 100107.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Smart stopword list.

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

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM.

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

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 international conference on World wide web*, pages 751–760. ACM.

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

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 783–792. ACM.

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 international ACM SIGIR conference on Research and development in information retrieval*, pages 323–332. ACM.

Peter Willett. 2006. The porter stemming algorithm: then and now. *Program*, 40(3):219–223.

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

Dingqi Yang, Daqing Zhang, Zhiyong Yu, and Zhu Wang. 2013. A sentiment-enhanced personalized location recommendation system. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 119–128. ACM.