Clustered Model Adaption for Personalized Sentiment Analysis

Lin Gong, Benjamin Haines, Hongning Wang
Department of Computer Science
University of Virginia, Charlottesville VA, 22904 USA
{lg5bt,bmh5wx,hw5x}@virginia.edu

ABSTRACT

We propose to capture humans' variable and idiosyncratic sentiment via building personalized sentiment classification models at a group level. Our solution roots in the *social comparison theory* that humans tend to form groups with others of similar minds and ability, and the *cognitive consistency theory* that mutual influence inside groups will eventually shape group norms and attitudes, with which group members will all shift to align. We formalize personalized sentiment classification as a multi-task learning problem. In particular, to exploit the clustering property of users' opinions, we impose a non-parametric Dirichlet Process prior over the personalized models, in which group members share the same customized sentiment model adapted from a global classifier. Extensive experimental evaluations on large collections of Amazon and Yelp reviews confirm the effectiveness of the proposed solution: it outperformed user-independent classification solutions, and several state-of-the-art model adaptation and multi-task learning algorithms.

CCS Concepts

•**Information systems** → **Sentiment analysis; Clustering and classification;**

Keywords

Sentiment analysis, model adaptation, multi-task learning

1. INTRODUCTION

Traditional solutions for text-based sentiment modeling mostly focus on building population-level supervised classifiers [29, 28, 36], which estimate and apply a shared classifier across all users' opinionated data. This postulates a strong assumption that the joint probability of sentiment labels and text content is independent and identical across users. However, this assumption is usually undermined in practice: it is well known in social psychology and linguistic studies that sentiment is personal and humans have diverse ways of expressing attitudes and opinions [37]. Hence, a single generic sentiment model can hardly capture the heterogeneity among users, and it will inevitably lead to inaccurate opinion mining results. Explicitly modeling the heterogeneity to capture individualized opinions is thus of particular importance.

Estimating a personalized sentiment model is challenging. Sparsity of individual users' opinionated data prevents us from estimating supervised classifiers on a per-user basis. Some existing works utilize semi-supervised methods to address the sparsity issue. For example, [18, 33] utilized user-user and user-document relations as regularizations to perform transductive learning. However, only one global sentiment model is estimated in such solutions, and it cannot capture the nuance in which individual users express their diverse opinions. [1] developed a transfer learning solution to adapt a global sentiment model to each individual user, but limited improvement is achieved on users with few observations, who form a major portion of the user population.

In this work, we take a new perspective to build personalized sentiment models by exploiting social psychology theories about humans' dispositional tendencies. First, the theory of social comparison [7] states that the drive for self-evaluation can lead people to associate with others of similar opinions and abilities, thus to form groups. This guarantees the relative homogeneity of opinions and abilities within groups. In our solution, we capture such clustering property of different users' opinions by postulating a non-parametric Dirichlet Process (DP) prior [12] over the individualized models, such that those models automatically form latent groups. In the posterior distribution of this postulated stochastic process, users join groups by comparing the likelihood of generating their own opinionated data in different groups (i.e., realizing self-evaluation and group comparison). Second, according to the cognitive consistency theory [25], once the groups are formed, members inside the same group will be influenced by other in-group members mutually through both implicit and explicit information sharing, which leads to the development of group norms and attitudes [32]. We formalize this by adapting a global sentiment model to individual users in each latent user group, and jointly estimating the global and group-wise sentiment models. The shared global model can be interpreted as the global social norm, because it is estimated based on observations from all users. It thus captures homogenous sentimental regularities across users. The group-wise adapted models capture heterogenous sentimental variations among users across groups. Because of this two-level information grouping and sharing, the complexity of preference learning will be largely reduced. This is of particular value for sentiment analysis in tail users, who only possess a handful of observations but take the major proportion in user population.

We should note that our notion of user group is different from those in traditional social network analysis, where user interaction or community structure is observed. In our solution, user groups are *latent*: they are formed based on the textual patterns in users'

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052693

tion parameter α in DP together with the the number of auxiliary variables m in sampling of $\{c_u\}_{u=1}^N$ play an important role in determining the number of latent user groups in all DP-based models. We empirically fixed $\alpha = 1.0$ and $m = 6$ in all such models. Due to the biased class distribution in both datasets, we compute F1 measure for both positive and negative class in each user, and take macro average among users to compare the different models' classification performance.

4.2 Feasibility of Automated User Grouping

First of all, it is important to verify our stochastic EM based posterior inference in cLinAdapt is converging, as only one sample was taken from the posterior of $\{c_u\}_{u=1}^N$ when updating the group sentiment models $\{\phi_c\}_{c=1}^\infty$ and global model θ^s. We traced the complete-data log-likelihood, the number of inferred latent user groups, together with the testing performance (by Eq (9)) during each iteration of posterior inference in cLinAdapt over all users from both datasets. We reported the results for the two datasets in Figure 2 and 3, where for visualization purpose the illustrated results were collected in every five iterations (i.e., thinning the sampling chain) after the burn-in period (the first ten iterations).

As observed from the results on both datasets, the likelihood kept increasing during the iterative posterior sampling process and converged later on. In the meanwhile, the group size fluctuated a lot at the beginning of sampling and became more stable near the end of iterations. On the other hand, the classification performance on the testing collection kept improving as more accurate sentiment models were estimated from the iterative sampling process. This verifies the effectiveness of our posterior inference procedure. We also looked into the automatically identified groups and found

Figure 4: Word clouds on Amazon.

Figure 5: Word clouds on Yelp.

many of them exhibited unique characteristics. The median number of reviews per user in these two datasets were only 7 and 8, while in some groups the average number of reviews per user is as large as 22.1, with small variances. This indicates active users were grouped together in cLinAdapt. In addition, the overall positive class ratio on these two datasets is 74.7% and 75.3% respectively, but in many identified groups the class distribution was extremely

Table 1: Effect of different feature groupings in cLinAdapt.

Method	Amazon		Yelp	
	Pos F1	Neg F1	Pos F1	Neg F1
Base	0.8092	0.4871	0.8809	0.6284
400-1600	0.8313	0.5033	0.8942	0.6563
400-all	0.8405	0.5213	0.8981	0.6632
800-1600	0.8325	0.5115	0.8959	0.6592
800-all	0.8437	**0.5478**	**0.9010**	**0.6694**
1600-all	**0.8440**	0.5334	0.8993	0.6674
all-all	0.8404	0.5391	0.8995	0.6681

biased: some towards negative, as low as 62.1% positive; and some towards positive, as high as 88.2% (note users with more than 90% positive or negative reviews have been removed). This suggests users with similar opinions were also successfully grouped in cLinAdapt. In addition, small fluctuation in the number of sampled user groups near the end of iterations is caused by a small number of users keeping switching groups (as new groups were created for them). This is expected and reasonable, since the group assignment is modeled as a random variable and multiple latent user groups might fit a user's opinionated data equally well. This provides us the flexility to capture the variance in different users' opinions.

In addition to the above quantitative measures, we also looked into the learnt word sentiment polarities reflected in each group's sentiment classifier to further investigate the automatically identified user groups. Most of the learnt feature weights followed our expectation of the words' sentiment polarities, and many words indeed exhibited distinct polarities across groups. We visualized the variance of learnt feature weights across all the groups using word clouds and demonstrated the top 10 words with largest variance and top 10 words with smallest variance in Figure 4 and 5 for Amazon and Yelp datasets respectively. Considering the automatically identified groups were associated with different number of users, we normalized the group feature weight vector by its $L2$ norm. The displayed size of the selected features in the word cloud is proportional to their variances. From the results we can find that, for example, the words "*bore*, *lack*, *worth*" conveyed quite different sentiment polarities among diverse latent user groups in Amazon dataset, while the words like "*pleasure, deal, fail*" had quite consistent polarities. This is also observed in the Yelp dataset, as we can find words like "*star*, *good*, *worth*" were used quite differently across groups, while the words like "*horribl*, *sick*, *love*" are used more consistently.

4.3 Effect of Feature Grouping

We then investigated the effect of feature grouping in cLinAdapt. As discussed in Section 3.3, different feature groupings can be applied to the individual models and global model, such that nonlinearity is introduced when different grouping functions are used in these two levels of model adaptation.

We adopted the most effective feature grouping method named "*cross*" from [35]. Following their design, we first evenly spilt the hold-out training set (for Base model training) into N non-overlapping folds, and estimated a single SVM model on each fold. Then, we created a $V \times N$ matrix by collecting the learned SVM weights from the N folds, on which k-means clustering was applied to group V features into K and L feature groups. We compared the performance of varied combinations of feature groups for individual and global models in cLinAdapt. 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 first column indicates the feature group sizes in the personal-

ized models and global model respectively. And *all* indicates one feature per group (i.e., no feature grouping). All adapted models in cLinAdapt achieved promising performance improvement against the Base model. In addition, further improved performance in cLinAdapt's was achieved when we increased the feature group size in the global model. Under a fixed feature group size in the global model, a moderate size of feature groups in personalized models was more advantageous.

These observations follow our expectation. Since the global model is shared across all users, the whole collection of training data can be leveraged to adapt the global model to overcome sparsity. This allows cLinAdapt to afford more feature groups in the global model, and leads to a more accurate model adaptation. But at the group level, data sparsity remains as the major bottleneck for accurate estimation of model parameters, although observations have already been shared in groups. Hence, the trade-off between observation sharing among features and estimation accuracy has to be made. Based on this analysis, we selected the combination of **800-all** feature grouping methods in the following experiments.

4.4 Personalized Sentiment Classification

We compared cLinAdapt against all nine baselines on both Amazon and Yelp datasets, and the detailed performance is reported in Table 2. Overall, cLinAdapt achieved the best performance against all baselines, except the prediction of positive class in Amazon dataset. Considering these two datasets are heavily biased towards positive class, improving the prediction accuracy in negative class is arguably more challenging and important.

It is meaningful to compare different algorithms' performance according to their model assumptions. First, as the Base model was trained on an isolated collection, though from the same domain, it failed to capture individual users' opinions. Global SVM benefited from gathering large collection of data from the targeted user population but was short of personalization, thus it performed well on positive class while suffered in negative class. Individual SVM could not capture each user's own sentiment model due to serious data sparsity issue; and it was the worst solution for personalized sentiment classification.

Second, as a state-of-the-art model adaptation based baseline, LinAdapt slightly improved over the Base model; but as the user models were trained independently, its performance was limited by the sparse observations in each individual user. The arbitrary user grouping by k-means barely helped LinAdapt in personalized classification, though more observations became available for model training. The joint user grouping with LinAdapt training finally achieved substantial performance improvement (especially on the Yelp dataset). Similar result was achieved in RegLR+DP as well. This confirms the necessity of joint task relatedness estimation and model training in multi-task learning.

Third, global information sharing is essential. All methods with a jointly estimated global model, i.e., MT-SVM, MT-RegLR+DP, cLinAdapt and also Global SVM, achieved significant improvement over others that do not have such a globally shared component. Additionally, as the class prior was against negative class in both datasets, observations of negative class became even rare in each user. As a result, compared with MT-SVM and MT-RegLR+DP baselines, cLinAdapt achieved improved performance in this class by sharing observations across features via its unique two-level feature grouping mechanism. However, comparing to MT-SVM, although no user grouping nor feature grouping was performed, its performance was very competitive. We hypothesized it was because on both datasets we had overly sufficient training signals for the globally shared model in MT-SVM. To verify this hypothesis, we reduced the number of users in the evaluation data set when training MT-SVM and cLinAdapt. Both models' performance decreased, but cLinAdapt decreased much slower than MT-SVM. When we only had five thousand users, cLinAdapt significantly outperformed MT-SVM in both classes on these two evaluation datasets. This result verifies our hypothesis and demonstrates the distinct advantage of cLinAdapt: when the total number of users (i.e., inductive learning tasks) is limited, properly grouping the users and leveraging information from a pre-trained model help improve overall classification performance.

Table 2: Personalized sentiment classification results.

Method	Amazon		Yelp	
	Pos F1	Neg F1	Pos F1	Neg F1
Base	0.8092	0.4871	0.8809	0.6284
Global SVM	0.8386	0.5245	0.8982	0.6596
Individual SVM	0.5582	0.2418	0.5691	0.3492
LinAdapt	0.8091	0.4894	0.8811	0.6281
LinAdapt+kMeans	0.8096	0.4990	0.8836	0.6461
LinAdapt+DP	0.8157	0.4721	0.8878	0.6391
RegLR+DP	0.8256	0.5021	0.8929	0.6528
MT-SVM	**0.8484**	0.5367	0.9002	0.6663
MT-RegLR+DP	0.8466	0.5247	0.8998	0.6630
cLinAdapt	0.8437	**0.5478**	**0.9010**	**0.6694**
Oracle-cLinAdapt	0.9049	0.6791	0.9268	0.7358

One limitation of cLinAdapt is that the latent group membership can only be inferred for users with at least one labeled training instance. This limits its application in cases where new users keep emerging for analysis. This difficulty is also known as cold-start, which concerns the issue that a system cannot draw any inferences for users about which it has not yet gathered sufficient information. One remedy is to acquire a few labeled instances from the testing users for cLinAdapt model update. But it would be prohibitively expensive if we do so for every testing user. Instead, we decide to only infer the group membership for the new users based on their disclosed labeled instances, while keep the previously trained cLinAdapt model intact (i.e., perform sampling defined in Eq (5) without changing the group structure). This implicitly assumes the previously identified user groups are comprehensive and the new users can be fully characterized by one of those groups.

In order to verify this testing scheme, we randomly selected 2,000 users with at least 4 reviews to create hold-out testing sets on both Amazon and Yelp reviews accordingly, and used the rest users to estimate the cLinAdapt model. During testing in each user, we held the first three reviews' labels as known, and gradually disclosed them to cLinAdapt to infer this user's group membership and classify in the rest reviews. For comparison purpose, we also included Individual SVM, LinAdapt and MT-SVM trained and tested in the same way on these two newly collected evaluation datasets for cold-start, and reported the results in Table 3. From the results, it is clear that Individual SVM's performance was almost random due to the limited amount of training data in this testing scenario. LinAdapt benefited from a predefined Base model, while the independent model adaptation in single users still led to suboptimal performance. The same reason also limited MT-SVM: it treats users independently by only sharing the global model among them, so that the newly available labeled instances could not effectively help individual models at beginning. cLinAdapt better handled cold-start by reusing the learned user groups for new users. Significant improvement was achieved for negative class, as the observations in negative class were even more scarce in those newly disclosed labeled instances of each testing user.

Table 3: Effectiveness of model sharing for cold-start on Amazon and Yelp.

Obs.	Amazon								Yelp							
	Individual SVM		LinAdapt		MT-SVM		cLinAdapt		Individual SVM		LinAdapt		MT-SVM		cLinAdapt	
	Pos F1	Neg F1	Pos F1	Neg F1	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}	0.0000	0.4203	0.8587	0.5898	0.8588	0.5073	**0.8925**	**0.6675**	0.0000	0.4101	0.9322	0.7724	0.9251	0.7285	**0.9582**	**0.8335**
2^{nd}	0.4683	0.3831	0.8455	0.5495	0.8534	0.5267	**0.8795**	**0.6076**	0.7402	0.3116	0.9243	0.7176	0.9291	0.7027	**0.9501**	**0.7726**
3^{rd}	0.7362	0.1751	0.8113	0.4863	0.8283	0.4919	**0.8440**	**0.5402**	0.7812	0.1608	0.8873	0.6639	0.8954	0.6619	**0.9116**	**0.7147**

Table 4: Collaborative filtering results on Amazon and Yelp.

Models	Amazon		Yelp	
	NDCG	MAP	NDCG	MAP
Average	0.7758	0.5587	0.6798	0.3867
LinAdapt	0.8046	0.6640	0.7445	0.4945
LinAdapt+kMeans	0.8030	0.6635	0.7399	0.4901
LinAdapt+DP	0.8004	0.6597	0.7454	0.4986
RegLR+DP	0.8023	0.6614	0.7460	0.4991
MT-SVM	0.8050	0.6646	0.7439	0.4935
MT-RegLR+DP	0.8030	0.6626	0.7419	0.4935
cLinAdapt	**0.8052**	**0.6652**	**0.7473***	**0.5001**

$^*p$-value<0.05 under binomal test

Another observation in Table 3 is that all models' testing performance decreased with more labeled instances disclosed from the testing users. This is unexpected and might indicate the consistence assumption about a user's sentiment model does not hold. To verify this, we tested an oracle setting of cLinAdapt in the original evaluation set: we revealed the labels of testing data when inferring group assignments in testing, and this greatly boosted the test performance of cLinAdapt. We appended the result in Table 2. This indicates the performance bottleneck of cLinAdapt is the accuracy of inferred group membership in testing phase. We assumed this membership is stationary in each user, but this might not be true given the reviews were generated in a chronological order and users' sentiment model might change over time. In our future work, we plan to also model the generation of document content in cLinAdapt, such that the inferred group membership can be calibrated for each testing document accordingly.

4.5 Serve for Collaborative Filtering

Collaborative filtering technique has been successfully applied in many recommendation systems. One of its key components is to infer the similarity between users. The learnt personalized sentiment model for each user naturally serves as a good proxy of their preference; and the distance between the model weights can therefore characterize the similarity between users. In this experiment, we evaluated the utility of learnt personalized models in collaborative filtering based recommendation. To create an evaluation data set, besides the items that a user has reviewed, we randomly selected a set of items from other users and label them as irrelevant in recommendation evaluation. We fixed this random item set to be four times large as a user's actually reviewed item size and maintained the same random candidate items in all the algorithms. In addition, we also removed the items that were only reviewed by one user. For each candidate item, we selected the target user's top K most similar neighbors who also reviewed this item, and calculated the weighted average of neighbors' actual ratings as ranking score for this item. Normalized discounted cumulative gain (NDCG) and mean average precision (MAP) are used to measure the recommendation quality. In particular, NDCG takes the user's original five star rating as a multi-scale relevance judgment, and MAP takes reviews with higher than 3 star as relevant and the rest as irrelevant.

We compared the recommendation performance based on the user similarity computed by different personalized sentiment classification methods on both Amazon and Yelp datasets. We also included a baseline that makes recommendations by the simple average of ratings from all the users who reviewed the item, and named it as Average. The average number of users who reviewed the same item in Amazon is 3.2 and in Yelp it is 10. Correspondingly, we selected top-4 neighbors and top-8 neighbors in Amazon and Yelp datasets respectively based on the cosine similarity between the users' personalized models. We report the resulting MAP and NDCG performance across all users in Table 4. As we can find from the Table 4, cLinAdapt achieved encouraging recommendation performance on both datasets, which indicates its learned sentiment models better captured the relatedness among users in their preferences over the recommended items. Despite the very sparse distribution of reviews in both datasets, cLinAdapt correctly recognized the preference of different users, and found the best neighbors for collaborative filtering.

5. CONCLUSION AND FUTURE WORK

In this paper, we developed a clustered model adaptation solution for personalized sentiment classification. Our work is inspired by the well-established social theories about humans' dispositional tendencies, i.e., social comparison and cognitive consistence. By exploiting the clustering property of users' sentiment models, empirically improved sentiment classification performance was achieved on two large collections of opinionated review documents.

Several areas are left open for our future explorations. In the current work, we assumed a user's latent group membership is stationary: once inferred from training data, it could be repeatedly used in testing. However, a user's group membership and even sentiment model might evolve over time. It is beneficial to efficiently update the model when new labeled data and users become available. Also, the current model is unable to inference group memberships over users with no labeled instances. This could be overcome if the generative model also accounts for the generation of review content in each user. In addition, it is interesting to study how to identify the feature grouping together with the user groups, such that the balance can be automatically adjusted with respect to the available training data in each latent user group.

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.

7. REFERENCES

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

[2] Charles E Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174, 1974.

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

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

[5] Jiang Bian, Xin Li, Fan Li, Zhaohui Zheng, and Hongyuan Zha. Ranking specialization for web search: a divide-and-conquer approach by using topical ranksvm. In *Proceedings of the 19th WWW*, pages 131–140. ACM, 2010.

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

[7] John Bruhn. The concept of social cohesion. In *The Group Effect*, pages 31–48. Springer, 2009.

[8] Meghana Deodhar and Joydeep Ghosh. Scoal: A framework for simultaneous co-clustering and learning from complex data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(3):11, 2010.

[9] Jean Diebolt and Eddie HS Ip. Stochastic em: method and application. In *Markov chain Monte Carlo in practice*, pages 259–273. Springer, 1996.

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

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

[12] Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[13] Leon Festinger. A theory of social comparison processes. *Human relations*, 7(2):117–140, 1954.

[14] Leon Festinger. *A theory of cognitive dissonance*, volume 2. Stanford university press, 1962.

[15] Bo Geng, Yichen Yang, Chao Xu, and Xian-Sheng Hua. Ranking model adaptation for domain-specific search. *TKDE*, 24(4):745–758, 2012.

[16] Giorgos Giannopoulos, Ulf Brefeld, Theodore Dalamagas, and Timos Sellis. Learning to rank user intent. In *Proceedings of the 20th CIKM*, pages 195–200. ACM, 2011.

[17] Lin Gong, Mohammad Al Boni, and Hongning Wang. Modeling social norms evolution for personalized sentiment classification.

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

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

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

[21] Yucheng Low, Deepak Agarwal, and Alexander J Smola. Multiple domain user personalization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 123–131. ACM, 2011.

[22] Julian McAuley, Rahul Pandey, and Jure Leskovec. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD*, pages 785–794. ACM, 2015.

[23] Brian Mullen and George R Goethals. *Theories of group behavior*. Springer Science & Business Media, 2012.

[24] Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[25] Theodore M Newcomb. *The acquaintance process.* Holt, Rinehart & Winston, 1961.

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

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

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

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

[30] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[31] Babak Shahbaba and Radford Neal. Nonlinear models using dirichlet process mixtures. *Journal of Machine Learning Research*, 10(Aug):1829–1850, 2009.

[32] Muzafer Sherif. *The psychology of social norms.* Harper, 1936.

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

[34] Sebastian Thrun and Joseph O'Sullivan. Discovering structure in multiple learning tasks: The tc algorithm. In *ICML*, volume 96, pages 489–497, 1996.

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

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

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

[38] Fangzhao Wu and Yongfeng Huang. Sentiment domain adaptation with multiple sources. In *Proceedings of the 54th Annual Meeting on Association for Computational Linguistics*, pages 301–310, 2016.

[39] Ya Xue, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, 8(Jan):35–63, 2007.

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