Cross Domain User Engagement Evaluation

Ali Montazeralghaem*, Hamed Zamani†, and Azadeh Shakery*

*School of Electrical and Computer Engineering, College of Engineering,
University of Tehran, Tehran, Iran
†Center for Intelligent Information Retrieval, University of Massachusetts Amherst,
Amherst, MA, USA
{ali.montazer,shakery}@ut.ac.ir,zamani@cs.umass.edu

Abstract. Due to the applications of user engagements in recommender systems, predicting user engagement has recently attracted considerable attention. In this task which is firstly proposed in ACM Recommender Systems Challenge 2014, the posts containing users' opinions about items (e.g., the tweets containing the users' ratings about movies in the IMDb website) are studied. In this paper, we focus on user engagement evaluation for cold-start web applications in the extreme case, when there is no training data available for the target web application. We propose an adaptive model based on transfer learning (TL) technique to train on the data from a web application and test on another one. We study the problem of detecting tweets with positive engagement, which is a highly imbalanced classification problem. Therefore, we modify the loss function of the employed transfer learning method to cope with imbalanced data. We evaluate our method using a dataset including the tweets of four popular and diverse data sources, i.e., IMDb, YouTube, Goodreads, and Pandora. The experimental results show that in some cases transfer learning can transfer knowledge among domains to improve the user engagement evaluation performance. We further analyze the results to figure out when transfer learning can help to improve the performance.

Keywords: User engagement, transfer learning, adaptive model, cold-start

1 Introduction

Twitter is a popular micro-blogging platform, which allows users to share their opinions and thoughts as fast as possible in very short texts. This makes Twitter a rich source of information with high speed of information diffusion. Therefore, several web applications (e.g., IMDb) have been integrated with Twitter to let people express their opinions about items (e.g., movie) in a popular social network [2, 10].

It is shown that the amount of users' interactions on tweets can be used to measure the users' satisfaction. In more detail, user engagements in Twitter has a strong positive correlation with the interest of users in the received tweets [2]. In addition, the purpose of recommender systems is to increase the satisfaction of users and thus, measuring the user engagements of tweets which contain the opinions of users about items (or products) can be employed to improve recommender systems performance [9].

In addition to recommender systems, user engagement evaluation has several other usages. For instance, Uysal and Croft [8] designed a personalized content filter based on user engagements in Twitter. Petrovic et al. [4] predicted whether a tweet will be retweeted or not. These works have focused on tweets with arbitrary content, while we are interested in engagement evaluation of tweets with predefined content[1].

Regarding the importance of user engagement evaluation in recommender systems, ACM Recommender Systems Challenge 2014[2] [6] has focused on ranking tweets of each user based on their engagements. This challenge only considered the tweets that are tweeted using the IMDb website. Similar to this challenge, in this paper the "*engagement*" value is computed as the total number of *retweets* and *favorites* that a tweet has achieved.

Recently, Zamani et al. [9] proposed an adaptive user engagement evaluation model for different web applications. They considered four popular web applications (also called domains) with wide variety of items. They proposed to employ multi-task learning to train a generalized model using all domains to improve the user engagement evaluation performance for each individual domain. Although their method successfully transfers knowledge among domains, it cannot be employed for evaluating user engagement for cold-start domains.

In this paper, we propose a cross domain adaptive model to train on one domain (source domain) and test on another one (target domain). In fact, the proposed method would be useful when there is no training data available for the target domain, i.e., cold-start web applications. To do so, we consider adaptive regularization-based transfer learning (ARTL) [3], which considers both distribution adaptation and label propagation strategies for cross domain transfer learning. Since distribution of our data is highly imbalanced[3], we modify the loss function of the ARTL method by adding an instance weighting term to the loss function formulation. To the best of our knowledge, this is the first try to evaluate user engagement in the case of absence of training data from the target domain.

In our experiments, we consider a collection of tweets from four popular web applications with very different items: IMDb (movie), YouTube (video clip), Pandora (music), and Goodreads (book) [9]. In our experiments, we analyze when transfer learning can help to improve the user engagement evaluation performance.

2 Cross Domain Model for User Engagement Evaluation

In this section, we first briefly explain the employed transfer learning algorithm and describe how we deal with imbalanced data in transfer learning scenarios. We further introduce our features for user engagement evaluation.

[1] In each tweets, the user gives a rate to or likes/dislikes a product.

[2] "User Engagement as Evaluation" Challenge, `http://2014.recsyschallenge.com/`

[3] There are lots of tweets with zero engagement and a few tweets with positive engagement.

2.1 Adaptive Regularization-based Transfer Learning

It is very difficult to induce a supervised classifier without any labeled data. Various transfer learning methods (also called domain adaptation methods) have been so far proposed to transfer knowledge from a source domain to a target domain, when there is no training data available for the target domain. In this paper, we employ adaptive regularization-based transfer learning (ARTL) [3], a cross domain transfer learning method whose goal is to improve classification performance for the unlabeled target domain using labeled data from the source domain.

Most existing transfer learning methods try to do one of the two following strategies: distribution adaptation and label propagation. ARTL framework considers both of these two strategies in its learning process. In fact, ARTL learns an adaptive classifier by optimizing the structural risk functional, the joint distribution matching between domains (J_s and J_t), and the manifold consistency underlying marginal distribution (P_s and P_t). Let $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ be a set of n training instances from the source domain in which $y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^d$ respectively denote the label and the feature vector, where d is the number of features. The ARTL framework is formulated as:

$$f = \arg\min_{f \in H_K} \mathcal{L}(f(X), Y) + \sigma \|f\|_K^2 + \lambda D_{f,K}(J_s, J_t) + \gamma M_{f,K}(P_s, P_t)$$

where K, H_K, $M_{f,K}$, $D_{f,K}$, and $\mathcal{L}$ respectively denote the kernel function, Hilbert space, manifold regularization, joint distribution adaptation, and the loss function. σ, λ, and γ are positive regularization parameters. Squared loss function is used in ARTL formulation.

Since the distribution of data in our problem is highly skewed, we propose to assign higher weights to instances from the minority class and vice versa. To this end, we define an instance weighting matrix $W \in R^{n \times 1}$ where elements of the matrix correspond to the weight of individual training instances. The matrix W is computed as:

$$W_i = \frac{1/n^{(i)}}{\sum_{j=1}^{n} 1/n^{(j)}}$$

where $n^{(i)}$ denotes the number of training instances with label y_i. A similar idea for coping with imbalanced data has been previously proposed in [1] for single-task classification and in [7, 9] for multi-task learning. We can now redefine the ARTL learning formulation as follows:

$$f = \arg\min_{f \in H_K} W\mathcal{L}(f(X), Y) + \sigma \|f\|_K^2 + \lambda D_{f,K}(J_s, J_t) + \gamma M_{f,k}(P_s, P_t)$$

2.2 Features

We extract 23 features from each tweet, that are partitioned into three categories: user-based, item-based, and tweet-based. Note that the contents of tweets in our task are predefined by the web applications and users usually do not edit tweets contents. These features are previously used in [9, 10]. More details about the exact definition of features can be found in [10]. The list of our features are as follows:

Table 1. Dataset Characteristics

	IMDb	**YouTube**	**Goodreads**	**Pandora**
# of tweets	100,206	239,751	65,445	98,212
# of users	6,852	6,480	3,813	3,312
# of items	13,502	154,041	31,558	32,321
Average engagement	0.1097	0.4737	0.1632	0.0778
% of tweets with positive engagement	4.139	14.193	6.931	6.285

User-based features. Number of followers, Number of followees, Number of tweets, Number of tweets about domain's items, Number of liked tweets, Number of lists, Tweeting frequency, Attracting followers frequency, Following frequency, Like frequency, Followers/Followees, Followers-Followees.
Item-based features. Number of tweets about the item.
Tweet-based features. Mention count, Number of hash-tags, Tweet age, Membership age at the tweeting time, Hour of tweet, Day of tweet, Time of tweet, Holidays or not, Same language or not, English or not.

3 Experiments

3.1 Experimental Setup

In our evaluations, we use the dataset provided by [9], which is gathered from four diverse and popular web applications (domains): IMDb, YouTube, Goodreads, and Pandora which contain movies, video clips, books, and musics, respectively.[4] Statistics of the dataset are reported in Table 1.

To have a complete and fair evaluation, in our experiments all models are trained using the same number of training instances. For each domain, we randomly select $16,361$ and $32,722$ instances to create training and test sets, respectively. We repeat this process 30 times using random shuffling. We report the average of the results obtained on these 30 shuffles and classify tweets with positive engagement from the tweets with zero engagement.

According to Table 1, the data is highly imbalanced; percentage of data with positive engagement is by far lower than percentage of those with zero engagement. In our evaluations, we consider accuracy (as the most popular evaluation metric for classification) and balanced accuracy (BA) [5] (a widely used evaluation metric for imbalanced situations). BA is computed as the arithmetic mean of accuracy in each class.

For single-task learning (STL), we employ support vector machine (SVM) classifier, which has been shown to be highly effective in various tasks. The linear kernel is considered for both baseline and the proposed method. To set the parameters of each learning algorithm, we perform hyper-parameter optimization using grid search and stratified k-fold ($k = 5$) cross validation. In addition, we apply instance weighting for both baseline and the proposed method in all the experiments.[5] We use the t-test with 95% confidence to capture the statis-

[4] The dataset is freely available at `http://ece.ut.ac.ir/node/100770`

[5] The results without instance weighting is biased toward the majority class. For the sake of space, the results without instance weighting are not reported.

Table 2. Accuracy and balanced accuracy achieved by single-task learning and transfer learning methods.

Train \ Test		IMDb		YouTube		Goodreads		Pandora	
		STL	ARTL	STL	ARTL	STL	ARTL	STL	ARTL
IMDb	BA	-	-	**0.6445***	0.6033	0.5802	**0.5911***	**0.5663***	0.5492
	Acc.			**0.7889***	0.6797	**0.8616***	0.6924	**0.8681***	0.6796
YouTube	BA	0.5378	**0.5542***	-	-	0.5534	**0.5582***	**0.5447***	0.5390
	Acc.	**0.9529***	0.9031			**0.9350***	0.9197	**0.9383***	0.9031
Goodreads	BA	0.5917	**0.5933**	**0.6767***	0.6506	-	-	**0.5752***	0.5572
	Acc.	**0.7830***	0.7008	0.5745	**0.6360***			**0.7557***	0.6720
Pandora	BA	0.5731	**0.5820***	**0.6602***	0.6368	0.5948	**0.5985**	-	-
	Acc.	**0.6835***	0.6525	0.5403	**0.6485***	**0.6769**	0.6682		

tically significant differences between results.

3.2 Results and Discussion

The results obtained by STL and ARTL are reported in Table 2. In this table, the significant differences between results are shown by star. According to this table, in some cases STL performs better and in other cases ARTL outperforms STL. In the following, we analyze the obtained results for each target domain.

IMDb. In the case that IMDb is the target domain, ARTL significantly outperforms STL, in terms of BA; however, the accuracy values achieved by SVM are higher than those obtained by ARTL. This shows that ARTL can classify the minority class instances (tweets with positive engagement) significantly better than SVM, but it fails in classifying the instances belonging to the majority class. The reason is that IMDb is the most imbalanced domain in the dataset (see Table 1) and thus, STL cannot learn a proper model, when there is a large gap between the feature distribution of the source and the target domains. This is why the maximum difference between the performance of ARTL and STL is happened when YouTube is selected as the source domain.

YouTube. Unlike the previous case, when YouTube is chosen as the target domain, STL performs better than ARTL in terms of BA. In some cases (i.e., training of Goodreads and Pandora) ARTL achieves higher accuracy compared to STL. The reason is that other domains are much more imbalanced compared to YouTube and in that case, the trained STL model is more accurate in detecting instances from the minority class, which leads to the better BA, but worse accuracy.

Goodreads. The results achieved over the Goodreads domain are very similar to those obtained over the IMDb domain. In other words, ARTL is more successful than STL in detecting tweets with positive engagement, since it achieved higher balanced accuracy but lower accuracy. As shown in Table 2, the best performance over this target domain is achieved when the model is trained using the IMDb or the Pandora domains. The percentage of data with positive engagement in these two domains are much more similar to Goodreads, compared to YouTube. Thus, learning from these domains can achieve higher accuracy.

Pandora. According to Table 2, transferring knowledge do not help to improve the user engagement evaluation performance. The reason could be related to the different distributions of the data from Pandora and the other domains. As reported in Table 1, the average engagement in this domain is much lower than the other domains which leads to have a very different feature distribution.

4 Conclusions and Future Work

In this paper, we proposed an adaptive method based on adaptive regularization-based transfer learning for user engagement evaluation. To cope with imbalanced data, we modified the transfer learning objective function by adding an instance weighting matrix to its formulation. In our experiments, we considered four popular web applications: IMDb, YouTube, GoodReads, and Pandora. The experimental results show that in some cases, we can find some useful information to transfer knowledge between these very different domains. We analyzed the achieved results and discussed the situations that transfer learning can be applied to improve the user engagement evaluation performance. An interesting future direction is to also modify the manifold regularization and the joint distribution adaptation components in the transfer learning objective function to improve the classification performance, when the data is highly imbalanced.

Acknowledgements. This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

1. Akbani, R., Kwek, S., Japkowicz, N.: Applying Support Vector Machines to Imbalanced Datasets. In: ECML. pp. 39–50 (2004)
2. Loiacono, D., Lommatzsch, A., Turrin, R.: An Analysis of the 2014 RecSys Challenge. In: RecSysChallenge. pp. 1–6 (2014)
3. Long, M., Wang, J., Ding, G., Pan, S.J., Yu, P.S.: Adaptation Regularization: A General Framework for Transfer Learning. IEEE Trans. Knowl. Data Eng. 26(5), 1076–1089 (2014)
4. Petrovic, S., Osborne, M., Lavrenko, V.: RT to Win! Predicting Message Propagation in Twitter. In: ICWSM. pp. 586–589 (2011)
5. Powers, D.: Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation. J. Mach. Learn. Tech. 2(1), 37–63 (2011)
6. Said, A., Dooms, S., Loni, B., Tikk, D.: Recommender Systems Challenge 2014. In: RecSys. pp. 387–388 (2014)
7. C. de Souza, J.G., Zamani, H., Negri, M., Turchi, M., Falavigna, D.: Multitask Learning for Adaptive Quality Estimation of Automatically Transcribed Utterances. In: NAACL-HLT. pp. 714–724 (2015)
8. Uysal, I., Croft, W.B.: User Oriented Tweet Ranking: A Filtering Approach to Microblogs. In: CIKM. pp. 2261–2264 (2011)
9. Zamani, H., Moradi, P., Shakery, A.: Adaptive User Engagement Evaluation via Multi-task Learning. In: SIGIR. pp. 1011–1014 (2015)
10. Zamani, H., Shakery, A., Moradi, P.: Regression and Learning to Rank Aggregation for User Engagement Evaluation. In: RecSysChallenge. pp. 29–34 (2014)