

# SODA : Service Oriented Domain Adaptation Architecture for Microblog Categorization

Himanshu S. Bhatt, Sandipan Dandapat, Peddamuthu Balaji, Shourya Roy,  
Sharmistha\* and Deepali Semwal\*

Xerox Research Centre India, Bangalore, INDIA

{Himanshu.Bhatt, Sandipan.Dandapat, Balaji.Peddamuthu2, Shourya.Roy}@xerox.com  
{sharmistha@ssl.serc.iisc.in, semwaldeepali@gmail.com}

## Abstract

We demonstrate SODA (Service Oriented Domain Adaptation) for efficient and scalable cross-domain microblog categorization which works on the principle of transfer learning. It is developed on a novel similarity-based iterative domain adaptation algorithm while extended with features such as active learning and interactive GUI to be used by business professionals. SODA demonstrates efficient classification accuracy on new collections while minimizing and sometimes eliminating the need for expensive data labeling efforts. SODA also implements an active learning (AL) technique to select informative instances from the new collection to seek annotations, if a small amount of labeled data is required by the adaptation algorithm.

## 1 Introduction

Online social media, such as Twitter.com, have become the de facto standard for sharing information, thoughts, ideas, personal feelings, daily happenings *etc.* which essentially led research and development in the field of social media analytics to flourish. Social media analytics provide actionable insights to business by analyzing huge amount of user generated content (UGC) (Sriram et al., 2010; Jo and Oh, 2011; He et al., 2012; Si et al., 2013; Nakov et al., 2013). Sentiment categorization, one of the common social media analytics task, segregates a collection of UGC into different buckets with positive, negative or neutral orientation (Liu and Zhang, 2012;

Thelwall et al., 2011; Bollen et al., 2009). This information is used to aggregate statistics and identify trends which are helpful for many applications viz. *Customer Care, Product Marketing, User Studies.*

Supervised machine learning (ML) techniques such as *text categorization* have played a key enabler role to classify microblogs into sentiment categories (Pang and Lee, 2008; Tan et al., 2009; Go et al., 2009; Fernández et al., 2014). These are trained on a fraction of annotated data as per client provided label set e.g.  $\{positive, negative, and neutral\}$  for a product/service/domain<sup>1</sup>. One of the obstacles towards rapid adoption of such systems is requirement of labeled tweets for developing ML-based models as it requires extensive human labeling efforts. Additionally, need of manual labeling slows down the process of categorization on high velocity social media which requires fast analytic insights. From our conversations with business professionals, we derived the need of a practical solution which would help them scale up across hundreds of collections and domains without the overhead of annotating data and building models from scratch every time for a new collection.

In this paper, we demonstrate *Service Oriented Domain Adaptation* (SODA) which offers social media analytics as-a-service to the users. Specifically, it provides sentiment categorization as-a-service that allows users to efficiently analyze comments from any new collection without the over-

<sup>1</sup>We use the word *collection* to describe tweets pertaining to a product/service/domain. For example tweets pertaining to *Apple iPhone Maps* and *Samsung S3 Battery* are two different collections. Tweets and comments are used interchangeably.

Work done at Xerox Research Centre India

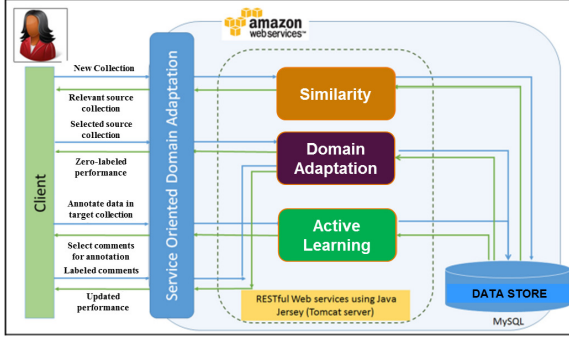


Figure 1: Overall architecture of SODA.

head of manual annotations or re-training models. It thus enables faster wide-scale analysis within and across different domains/industries such as telecom, healthcare, finance etc. SODA is based on an iterative ensemble based adaptation technique (Bhatt et al., 2015) which gradually transfers knowledge from the source to the new target collection while being cognizant of similarity between the two collections. It has been extensively evaluated by business professionals in a user-trial and on a benchmark dataset.

## 2 SODA Features

Figure 1 illustrates the architecture of SODA comprising three primary modules, 1) similarity, 2) domain adaptation, and 3) active learning. The first two modules use unlabeled data from the new collection while the optional third module helps in creating labeled data for enhanced classification performance. These modules are explained below.

### 2.1 Similarity

In social media analytics, especially for sentiment categorization, there exist numerous collections about different products or services where labeled data is available and thus can be used to adapt to a new unlabeled collection. Given a target collection, the key question is to identify the best possible source collection to adapt from. The similarity module in SODA identifies the best adaptable source collection based on the similarity between the source and target collections. This is based on the observations from existing literature (Bhatt et al., 2015; Blitzer et al., 2007) which suggest that if the source and target collections are similar, the adaptation performance tends to be better than if the two collections are dissimilar. The similarity module in SODA is capable of computing different kinds of

lexical, syntactic, and semantic similarities between unlabeled target and labeled source collections. For this demonstration on sentiment categorization from social media data, it measures cosine similarity between the comments in each collection and computes  $sim$  as the similarity score.

### 2.2 Domain Adaptation

The heart of SODA is the adaptation module that works on two principles, *generalization* and *adaptation*. During generalization, it learns shared common representation (Blitzer et al., 2007; Ji et al., 2011; Pan et al., 2010) which minimizes the divergence between two collections. We leverage one of the widely used structural correspondence learning (SCL) approach (Blitzer et al., 2007) to compute shared representations. The idea adhered here is that a model learned on the shared feature representation using labeled data from the source collection will also generalize well on the target collection. Towards this, we learn a model ( $C_S$ ) on the shared feature representation from the source collection, referred to as “source classifier”.  $C_S$  is then used to predict labels for the pool of unlabeled instances from the target collection, referred to as  $P_u$ , using the shared representations. All instances in  $P_u$  which are predicted with a confidence ( $\alpha_1$ ) higher than a predefined threshold ( $\theta_1$ ) are moved to the pool of pseudo-labeled target instances, referred to as  $P_s$ . We now learn a target domain model  $C_T$  on  $P_s$  using the target specific representation, referred to as “target classifier”.

$C_T$  captures a separate view of the target instances than the shared representation and hence brings in discriminating target specific information which is useful for categorization in target collection. For further adaptation, the source ( $C_S$ ) and target ( $C_T$ ) classifiers are combined in a weighted ensemble ( $E$ ) with  $w_s$  and  $w_t$  as the corresponding weights and iterate over the remaining unlabeled instances in  $P_u$ . In each iteration, the ensemble processes the remaining instances and iteratively adds confidently predicted instances to  $P_s$  which are used to re-train/update  $C_T$ . This iterative process continues till all instances in  $P_u$  are confidently labeled or a maximum number of iterations is reached. Transfer occurs within the ensemble where the source classifier progressively facilitates the learning of tar-

## Algorithm 1 Adaptation Algorithm

**Input:**  $C_s, Q, C_t, P_s, \& P_u$ .  
**Iterate:**  $l = 0$  : till  $P_u = \{\phi\}$   
**Process:** Construct ensemble  $E \rightarrow w_l^s C_s + w_l^t C_t$ . Initialize  $w_l^s$  &  $w_l^t$  to 0.5.  
**for**  $i = 1$  to  $n$  (size of  $P_u$ ) **do**  
    Predict labels:  $E(Q\mathbf{x}_i, \mathbf{x}_i) \rightarrow \hat{y}_i$ ; confidence of prediction:  $\alpha_i$ .  
    **if**  $\alpha_i > \theta_2$  **then**  
        Move  $i^{th}$  instance from  $P_u$  to  $P_s$  with pseudo label  $\hat{y}_i$ .  
    **end if.**  
**end for.**  
Re-train  $C_t$  & update  $w_l^s$  and  $w_l^t$  as:  

$$w_{(l+1)}^s = \frac{(sim * w_l^s * I(C_s))}{(sim * w_l^s * I(C_s) + (1 - sim) * w_l^t * I(C_t))}$$

$$w_{(l+1)}^t = \frac{((1 - sim) * w_l^t * I(C_t))}{(sim * w_l^s * I(C_s) + (1 - sim) * w_l^t * I(C_t))}$$
**end iterate.**  
**Output:** Final  $C_t$  and updated weights  $w^s$  and  $w^t$

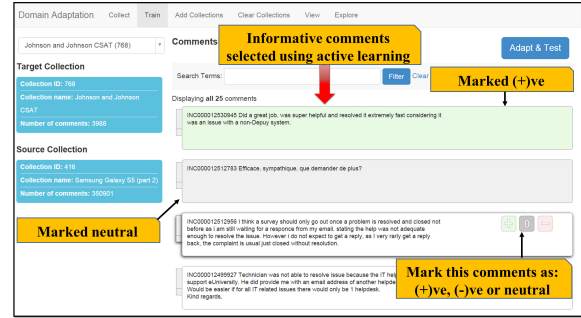
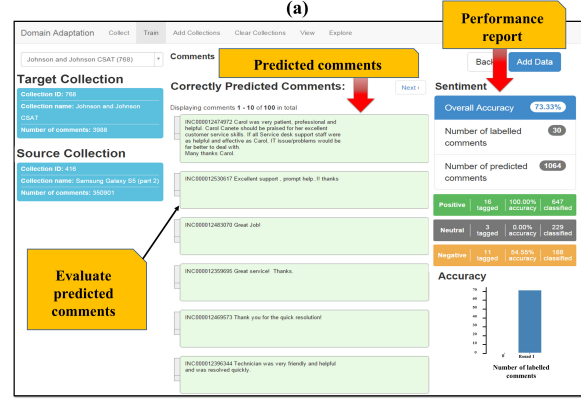
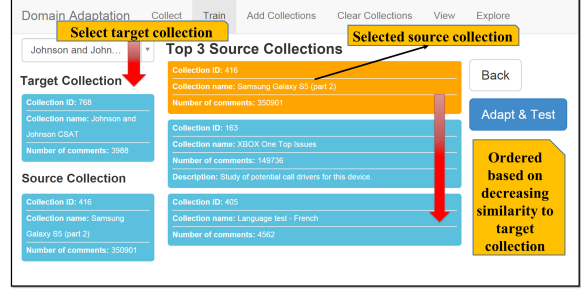
get classifier by providing pseudo labeled instances. The weights of the individual classifiers are updated, as a function of error ( $I(\cdot)$ ) and the similarity ( $sim$ ) between the collections, which gradually shift the emphasis from source to the target classifier. Finally, the ensemble is used to predict labels for future unseen instances in the target collection. Algorithm 1 summarizes our approach (refer (Bhatt et al., 2015) for more details).

## 2.3 Active Learning

SODA also implements an active learning module to allow users to annotate a few selected informative comments from the target collection. These comments are selected using cross entropy difference (CED) (Axelrod et al., 2011) such that the difference with source collection and the similarity with target collection is maximized. It selects comment(s) from target collection that have low CED score i.e. comments that have high entropy with respect to source  $H_S(\cdot)$  and low entropy with respect to target collection  $H_T(\cdot)$  as in Equation (1).

$$CED(s) = H_T(s) - H_S(s) \quad (1)$$

Note, this active learning module is optional and should be used when the adaptation performance with unlabeled instances is not satisfactory. More and more instances can be annotated in multiple rounds till a satisfactory performance is achieved. These annotated instances are used to build a stronger target classifier for the ensemble based adaptation algorithm.



(c)

Figure 2: Interactive GUI of service oriented domain adaptation (SODA) (best viewed in color).

## 3 Design and Internals

Figure 2(a) illustrates the interactive user interface (UI) of SODA where one can select a new target collection for the analysis task (i.e. sentiment categorization). For a new target collection, it identifies relevant adaptable source collections based on their similarity. One can select any of the candidate source collections (selected collection highlighted in Figure 2(a)) and adapt. Figure 2(b) shows the performance report along with the predicted comments from the target collection. User evaluates the adaptation performance in unlabeled target collection by analyzing the predicted comments and decides whether to annotate additional comments in

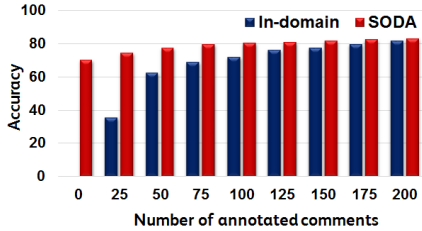


Figure 3: The effect of labeled comments on the performance while adapting from Coll-1  $\rightarrow$  Coll-6.

target collection? If yes, Figure 2(c) lists a few informative comments selected using the active learning module to seek annotations. One can mark these comments as positive, negative or neutral and subsequently adapt using these labeled instances from the target collection. Figure 2(b) also shows the adaptation performance with a few labeled instances in the target collection. One can continue annotating more instances in the target collection until satisfactory performance is achieved. For more detailed demonstration, please refer to the video.<sup>2</sup>

The interactive UI of SODA is developed using Ruby on Rails framework. All collections are managed in MySQL server. All three modules in SODA fetch data from the server and write the output back to the server. All modules work in real time enabling the system to be highly responsive to the user. The application is hosted on Amazon AWS as RESTful web services using Java Jersey (Tomcat server) that act as a bridge between the UI and back end.

#### 4 User Trial & Experimental Results

To evaluate the overall experience, a user trial was conducted where several business professionals provided feedback on SODA. The objective was to evaluate the overall usability, reduction in required efforts, and the performance on the new target collections. The overall evaluation rated SODA 5 on usability and 4 for reduction in efforts (1 being worst & 5 being the best). Table 1 reports the classification accuracy of SODA with few labeled comments from the target collection (ranging from 0 to 100). It also reports the performance of the in-domain classifier which is trained and tested on data from the same collection. *Coll-1* to *Coll-8* refer to collections pertaining to *marketing & sales, comcast support,*

<sup>2</sup><https://www.youtube.com/watch?v=zKnP5QEHVAE>

Table 1: User-trial results on social media data.

Source	Target	Accuracy with annotations					In-domain
		0	25	50	75	100	
Coll-1	Coll-6	70	74	77	79	<b>80</b>	71.7
Col-2	Coll-8	73	45	71	73	<b>79</b>	74.5
Coll-3	Coll-5	87	94	94	93	<b>95</b>	92.0
Coll-4	Coll-7	83	84	94	96	<b>96</b>	85.7

Table 2: Results on the Amazon review dataset.

Target	Source	BL	SCL	SODA	In-domain
B	E	61.5	76.1	78.9	<b>80.4</b>
	K	68.4	66.2	74.8	
	D	60.1	77.6	80.0	
E	B	68.2	77.8	80.3	84.4
	D	61.5	74.1	76.4	
	K	76.2	83.1	<b>85.2</b>	
K	B	71.7	78.8	80.1	87.7
	D	60.3	79.4	82.0	
	E	73.2	83.8	<b>87.9</b>	
D	B	63.5	74.2	81.9	<b>82.4</b>
	E	62.3	74.8	80.1	
	K	67.4	76.5	78.8	

*DirectTV support, ASUS, Johnson & Johnson CSAT, Apple iPhone6, and HUWAEI* respectively. Figure 3 compares the effect of adding labeled comments in batches of 25 comments at-a-time. When there is no labeled data in the target collection, in-domain classifier can not be applied while SODA still yields good classification accuracy. Moreover, SODA consistently performs better than the in-domain classifier with same amount of labeled data.

We also evaluated the performance of domain adaptation (DA) module of SODA on the Amazon review dataset (Blitzer et al., 2007) which is a benchmark dataset for sentiment categorization. It has 4 domains, namely, books(**B**), dvds(**D**), electronics(**E**), and kitchen(**K**) each with 2000 reviews divided equally into positive and negative reviews. Table 2 shows that DA module of SODA outperforms 1) a widely used domain adaptation technique, namely, structural correspondence learning (SCL) (Blitzer et al., 2007; Blitzer et al., 2006), 2) the baseline (**BL**) where a classifier trained on one domain is applied on another domain, and 3) the in-domain classifier. Note that in Table 2, the performance of DA module of SODA is reported when it does not use any labeled instances from the target domain.

#### 5 Conclusion

We demonstrated SODA for efficient microblog categorization on new social media collections with minimum (or sometimes no) need of manual annotations; thus, enabling faster and efficient analytics.

## References

- Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 355–362.
- Himanshu Sharad Bhatt, Deepali Semwal, and Shourya Roy. 2015. An iterative similarity based adaptation technique for cross-domain text classification. In *Proceedings of Conference on Computational Natural Language Learning*.
- John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 120–128.
- John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of Association for Computational Linguistics*, pages 187–205.
- Johan Bollen, Alberto Pepe, and Huina Mao. 2009. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *CoRR*.
- Javi Fernández, Yoan Gutiérrez, José M Gómez, and Patricio Martínez-Barco. 2014. Gplsi: Supervised sentiment analysis in twitter using skipgrams. In *Proceedings of International Workshop on Semantic Evaluation (SemEval)*, pages 294–299.
- Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1:12.
- Yulan He, Chenghua Lin, Wei Gao, and Kam-Fai Wong. 2012. Tracking sentiment and topic dynamics from social media. In *Proceedings of International AAAI Conference on Weblogs and Social Media*.
- Yang-Sheng Ji, Jia-Jun Chen, Gang Niu, Lin Shang, and Xin-Yu Dai. 2011. Transfer learning via multi-view principal component analysis. *Journal of Computer Science and Technology*, 26(1):81–98.
- Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of International Conference on Web Search and Data Mining*, pages 815–824.
- Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining Text Data*, pages 415–463.
- Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter.
- Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings International Conference on World Wide Web*, pages 751–760.
- Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.
- Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. 2013. Exploiting topic based twitter sentiment for stock prediction. In *Proceedings of Association for Computational Linguistics*, volume 2013, pages 24–29.
- Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short text classification in twitter to improve information filtering. In *Proceedings of International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 841–842.
- Songbo Tan, Xueqi Cheng, Yuefen Wang, and Hongbo Xu. 2009. Adapting naive bayes to domain adaptation for sentiment analysis. In *Advances in Information Retrieval*, pages 337–349.
- Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2011. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418.