Feature Ensemble Plus Sample Selection: Domain Adaptation for Sentiment Classification (Extended Abstract) *

Rui Xia †[1,3], Chengqing Zong[2], Xuelei Hu[1] and Erik Cambria[4]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
[3]State Key Laboratory of Novel Software Technology, Nanjing University, China
[4]School of Computer Engineering, Nanyang Technological University, Singapore

Abstract

The domain adaptation problem arises often in the field of sentiment classification. There are two distinct needs in domain adaptation, namely labeling adaptation and instance adaptation. Most of current research focuses on the former one, while neglects the latter one. In this work, we propose a joint approach, named feature ensemble plus sample selection (SS-FE), which takes both types of adaptation into account. A feature ensemble (FE) model is first proposed to learn a new labeling function in a feature re-weighting manner. Furthermore, a PCA-based sample selection (PCA-SS) method is proposed as an aid to FE for instance adaptation. Experimental results show that the proposed SS-FE approach could gain significant improvements, compared to individual FE and PCA-SS, due to its comprehensive consideration of both labeling adaptation and instance adaptation.

1 Introduction

The problem of domain adaptation has attracted increasing attention in the fields of both machine learning and natural language processing (NLP). Domain adaptation arises often in sentiment classification [Cambria *et al.*, 2014]. For example, we want to build a book review classifier, in case that the labeled book reviews in hand are scarce but the labeled movie reviews are abundant. Therefore, we need to adapt a classifier trained by labeled reviews in the movie domain to the book domain. Researchers have proposed a variety of domain adaptation approaches in the literatures. In general, the goal of domain adaptation is to build a machine learning model that maximizes the target-domain joint likelihood based on the source-domain labeled data:

$$\theta^* = \arg\max_{\theta} \int_{\mathbf{x}} \sum\nolimits_{y} p_t(\mathbf{x}, y) \log p(\mathbf{x}, y|\theta) d\mathbf{x}$$
$$\approx \arg\max_{\theta} \int_{\mathbf{x}} \tilde{p}_s(\mathbf{x}) \sum\nolimits_{y} \tilde{p}_s(y|\mathbf{x}) \log p(\mathbf{x}, y|\theta) d\mathbf{x},$$

where $p(\mathbf{x}, y|\theta)$ denotes the joint distribution of instance $\mathbf{x}$ and class label y defined in the learning algorithm. $\tilde{p}_s(\mathbf{x})$ and $\tilde{p}_s(y|\mathbf{x})$ denote the approximated target-domain distribution, which should be estimated by the source-domain labeled data.

According to the analysis in [Jiang and Zhai, 2007], there are two distinct needs in domain adaptation: *instance adaptation* and *labeling adaptation*. We interpret the two needs as follows:

- **Labeling adaptation** models the adaptation process of the labeling function $\tilde{p}_s(y|\mathbf{x}) \rightarrow p_t(y|\mathbf{x})$. Since one term that is positive in the source domain might express an opposite sentiment in the target domain, labeling adaptation aims to learn a new labeling function (or feature representation) for the target domain, using source domain-labeled data as well as a small amount of labeled (or a large amount of unlabeled) target-domain data;
- **Instance adaptation** models the adaptation process of instance distribution $\tilde{p}_s(\mathbf{x}) \rightarrow p_t(\mathbf{x})$. Since different domains have different term frequencies, instance adaptation aims to approximate the target-domain distribution by assigning different weights to the source-domain labeled data and then conducting importance sampling.

[Pan and Yang, 2010] presented a survey on transfer learning, which categorizes transfer learning approaches in a similar manner: feature-based transfer, instance-based transfer, and model-parameter-based transfer.

Existing work for domain adaptation in sentiment classification mostly belongs to labeling adaptation. [Blitzer *et al.*, 2007] proposed the structural correspondence learning (SCL) algorithm. [Daumé III, 2007] introduced a simple method for domain adaptation based on feature space augmentation.

*This paper is an extended abstract of the journal publication [Xia *et al*., 2013b].

†Email: rxia@njust.edu.cn

Pan proposed two domain adaptation approaches via transfer component analysis (TCA) [Pan *et al.*, 2010] and spectral feature alignment (SFA) [Pan *et al.*, 2011] , respectively. [Xia and Zong, 2011] and [Samdani and Yih, 2011] proposed feature reweighing based on an ensemble of feature sets, although they divide the feature sets in different ways. [Cambria and Hussain, 2015] employed PCA and linear discriminant analysis to build SenticNet, a concept-level resource for open-domain sentiment analysis. They also used affective knowledge to adapt general-purpose common-sense knowledge to the task of polarity detection [Cambria *et al.*, 2015].

The instance adaptation problem is also known as *sample selection bias* in the machine learning scenario. To address this problem, [Sugiyama *et al.*, 2008] proposed a Kullback-Leibler importance estimation procedure (KLIEP) algorithm.

In the field of NLP, [Axelrod *et al.*, 2011] proposed a method called pseudo in-domain data selection to select source domain training samples based on a language model, for cross-domain machine translation. [Xia *et al.*, 2013a] proposed an instance selection and instance weighting approach via PU learning (PUIS and PUIW) for the task of cross-domain sentiment classification. However, current researches in domain adaptation focused on either labeling adaptation or instance adaptation, individually. To the best of our knowledge, research on modeling two kinds of adaptation together is pretty scarce. In this work, we propose a joint method, called feature ensemble plus sample selection (SS-FE), to take full account of these two attributes for domain adaptation in sentiment classification. This approach could yield significant improvements compared to individual feature ensemble (FE) or sample selection (SS) methods, because it comprehensively considers both labeling adaptation and instance adaptation. For more details of this work, see [Xia *et al.*, 2013b].

2 The Proposed Approach

2.1 Feature Ensemble

In formulating our SS-FE method, we first propose a labeling adaptation method via POS-based feature ensemble (FE). This idea is based on the observation that, features with different type of POS tags have a distinct change in distribution in domain adaptation. We term the heavily changing features *domain-specific*, and the slightly changing features *domain-independent*. The domain-independent features generally perform more consistently when the domain changes. E.g., some adjectives and adverbs, such as "great" and "like," always have a strong correlation with the positive class label, regardless of domain. However, the domain-specific features may indicate different sentiment in different domains [Cambria *et al.*, 2013], e.g., in the concept "go read the book," the noun "book" most likely indicates a positive sentiment for book reviews, a but negative sentiment for movie reviews.

POS tags are supposed to be significant indicators of sentiment. Previous work revealed a high correlation between the presence of adjectives and document sentiment; certain verbs and nouns are also strong indicators of sentiment [Liu, 2012]. For cross-domain sentiment classification, we observe that features with different types of POS tags might have different levels of distributional change in domain adaptation. For example, nouns change the most because domains are mostly indicated by nouns, while adjectives and adverbs are fairly consistent across domains. The cross-domain Kullback-Leibler (K-L) distance regarding different POS tags (reported in the section of Experiments) confirms our observation.

Based on this observation, we divide features into four groups: adjectives and adverbs (J), verbs (V), nouns (N), and the others (O). Base-classifiers will be trained on all four feature subsets.Thus, a feature vector x is made up of four parts: $\mathbf{x} = [\mathbf{x}_J, \mathbf{x}_V, \mathbf{x}_N, \mathbf{x}_O]$, and we use $g_k(\mathbf{x}_k)$ to denote the labeling function of each base-classifier:

$$g_k(\mathbf{x}_k) = w_k^T \mathbf{x}_k, k \in \{J, V, N, O\}.$$

After base classification, we use the Stacking algorithm for meta-learning, where we construct the meta-learning feature vector $\tilde{\mathbf{x}} = [g_J, g_V, g_N, g_O]$ on a small amount of labeled data from the target domain (the validation set), and the weights of base-classifiers are optimized by minimizing the perceptron loss so that each components final weights are tuned to adapt to the target domain. We represent the weighted ensemble as

$$f(\mathbf{x}) = \sum_k \theta_k g_k(\mathbf{x}_k) == \sum_k \theta_k \sum_i w_{ki} x_i,$$

where θ_k is the ensemble weight. In the meta-learning process, we expect the algorithm to assign larger weights to the domain-independent parts, such as adjectives and adverbs, and lower weights to the domain-specific parts, such as nouns.

2.2 Sample Selection

The FE approach adapts the labeling function $\tilde{p}_s(y|\mathbf{x}) \rightarrow p_t(y|\mathbf{x})$ (in our approach, $g \rightarrow f$). However, the labeling adaptation is based on the condition that the instance probability $p_s(\mathbf{x})$ and $p_t(\mathbf{x})$ are the same. If there is a big gap between them, the effect of labeling adaptation will be reduced. To address this issue, we propose PCA-SS as an aid to FE. PCA-SS first selects a subset of the source domain labeled data whose instance distribution is close to the target domain, and then uses these selected samples as training data in labeling adaptation. Let $\mathbf{X}_t$ denote the document-by-feature matrix of the target domain dataset; we then get the target-domains latent concepts by solving the following SVD problem:

$$\mathbf{X}_t = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T,$$

where $\mathbf{U}$ and $\mathbf{V}$ are unit orthogonal matrices and $\mathbf{\Sigma}$ contains the positive singular values of decreasing magnitude $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_M$. The latent concepts are the orthogonal column vectors in the matrix $\mathbf{V}$, and the variance of the data projected along the i-th column of $\mathbf{V}$ is equal to σ_i^2.

To optimally capture the datas variations, only those latent concepts corresponding to the $k (k < M)$ largest singular values are typically retained. By selecting the columns of $\mathbf{P} = \mathbf{V}[:, 0 : k] \in \mathbf{R}^{M*k}$, which correspond to the latent concepts associated with the first k singular values as the projection matrix, we obtain the document-by-concept matrix

$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{P}.$$

Note that Hotelling's T^2 statistic reflects the degree of the magnitude deviation of each sample in a PCA model. In the settings of domain adaptation, it measures the extent to which a sample deviates from the concept space. Therefore, as the criterion for sample selection, we consider the T^2 statistic as a measure of "concept distance":

$$D(\mathbf{x}) = T^2(\mathbf{x}) = \mathbf{z}^T\mathbf{z} = \mathbf{x}\mathbf{P}\mathbf{\Lambda}_k^{-1}\mathbf{P}^T\mathbf{x}^T,$$

where $\mathbf{z} = \mathbf{\Lambda}_k^{-\frac{1}{2}}\mathbf{P}^T\mathbf{x}^T$, and $\mathbf{\Lambda}_k$ is the diagonal matrix corresponding to the top k singular values.

By this definition, we can obtain a concept distance for each sample $x_s^{(n)}$ in the source domain:

$$D(\mathbf{x}_s^{(n)}) = \mathbf{x}_s^{(n)}\mathbf{P}\mathbf{\Lambda}_k^{-1}\mathbf{P}^T{\mathbf{x}_s^{(n)}}^T.$$

Correspondingly, we get a set of the concept distances for each sample $\mathbf{x}_t^{(n)}$ in the target domain: $D(\mathbf{x}_t^{(n)}) = \mathbf{x}_t^{(n)}\mathbf{P}\mathbf{\Lambda}_k^{-1}\mathbf{P}^T{\mathbf{x}_t^{(n)}}^T$. Finally, we define the sample selection threshold as:

$$\bar{D} = \max_{x_t^{(n)} \in \mathbf{X}_t} \{D(\mathbf{x}_t^{(n)})\}.$$

Samples in the source domain $\mathbf{x}_s^{(n)}$ with $D(\mathbf{x}_s^{(n)}) > \bar{D}$ are discarded, and those with lower concept distance than the threshold are selected as training samples in domain adaptation.

2.3 Feature Ensemble Plus Sample Selection

In the above-mentioned techniques, FE models only labeling adaptation and neglects instance adaptation, while PCA-SS considers only instance adaptation. Therefore, we combine PCA-SS and FE in a serial manner, and refer to the joint approach as SS-FE, to address two types of adaptation together.

we first employ PCA-SS to project the data onto a latent target-domain concept space, select a subset of source-domain samples that are close to the target domain. We then apply FE to the selected part of data set. In FE, we first train individual classifiers with different feature sets divided by their POS tags. The final model is a weighted ensemble of individual classifiers, in which the weights are turned with the goal of increasing the weight of domain-independent features and reducing the weight of domain-specific features.

We empirically show that both FE and PCA-SS are effective for cross-domain sentiment classification, and that SS-FE performs better than either approach because it comprehensively considers both labeling and instance adaptation.

3 Experiments

3.1 Datasets and Experimental Setup

We use the multi-domain sentiment datasets [Daumé III, 2007] for experiments. It consists of product reviews collected from four different domains of Amazon.com: book (B), DVD (D), electronics (E), and kitchen (K). Each domain contains 1,000 positive and 1,000 negative reviews. The term "source → target" is used to denote different cross-domain tasks. For example, "D → B" represents the task that is trained in the DVD domain but tested in the book domain.

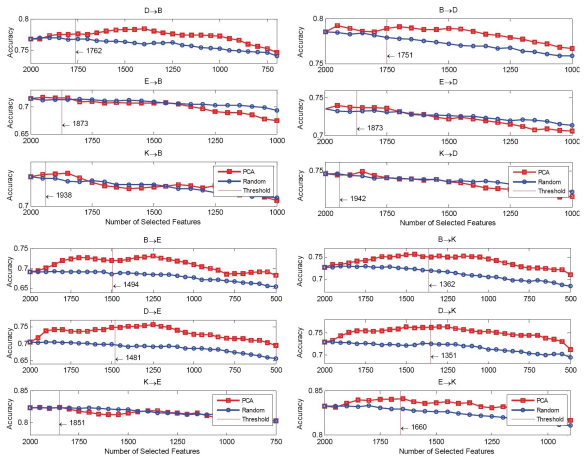

Figure 1: The performance of principal component analysis-based sample selection (PCA-SS)

In each of the tasks, labeled instances in the source domain are used for training base-classifiers. Labeled instances in the target domain are evenly split into 10 folds, where one fold is used as a validation set for meta-learning, and the other nine folds are used as a test set. All of the following experimental results are reported in terms of an average of the 10 folds' cross validation.

3.2 Experimental Results of PCA-SS

We first present the results of sample selection. The number of principal components is chosen with the percentage of variance contribution larger than 99.5 percent. All samples in the source and target domains are projected onto the concept space. We sort all source domain samples in an ascending order, according to their concept distance, and present the classification accuracy trained with a decreasing number of selected features in Figure 1. We chose naïve Bayes (NB) as the base classification algorithm, as it reportedly performs the best among three classifiers (NB, MaxEnt, and SVM) in the multi-domain dataset [Xia *et al.*, 2011]. For comparison, we also report the performance of the same number of randomly selected samples. We observe the experimental results from the following perspectives.

Random selection. We first observe the result of randomly selected samples. In all tasks, performance is gradually decreased by reducing the number of selected samples. This is in accordance with our standard understanding that decreasing training samples hurts machine learning performance.

PCA-based sample selection. In most of the tasks, the performance increases when the number of selected samples decreases at the first stage. After reaching the peak value, the performance gradually decreases. This leads to the conclusion that using a selected subset of samples as training data may improve classification performance over training on all samples. The conclusion holds well in the tasks of D → B, B → D, B → E, D → E, B → K, D → K. It is also worth noticing that sample selection is not quite effective in some tasks (such as E → B, K → D, and K → E). We will discuss why later.

Table 1: Cross-domain classification accuracy.

Tasks	Base-J	Base-V	Base-N	Base-O	Unigrams	FE	PCA-SS	SS-FE
D→B	0.7589	0.668	0.6307	0.6432	0.7685	0.7987	0.7759	**0.8038**
E→B	0.6907	0.6015	0.6038	0.5938	0.7156	0.7163	0.7162	**0.7286**
K→B	0.7034	0.5997	0.5943	0.5974	0.7265	**0.7334**	0.7285	0.7294
B→D	0.7665	0.6871	0.6597	0.6348	0.7854	0.7874	0.7889	**0.7910**
E→D	0.7228	0.5877	0.6142	0.5743	0.7350	0.7296	0.7373	**0.7460**
K→D	0.7179	0.5994	0.6142	0.5766	0.7469	0.7563	0.7458	**0.7570**
B→E	0.7319	0.6186	0.5649	0.5755	0.6915	0.7223	0.7199	**0.7424**
D→E	0.7431	0.6168	0.6005	0.5624	0.7040	0.7496	0.7477	**0.7707**
K→E	0.8051	0.7185	0.6961	0.6076	0.8235	0.8226	0.8243	**0.8293**
B→K	0.7568	0.6328	0.5936	0.5916	0.7265	0.7412	0.7507	**0.7807**
D→K	0.7334	0.6228	0.5987	0.6025	0.7299	0.7644	0.7630	**0.7782**
E→K	0.8083	0.7285	0.7037	0.6237	0.8330	0.8324	0.8412	**0.8487**

3.3 Experimental Results of SS-FE

In this section, we present the results of SS-FE. Four base NB classifiers are trained on the four feature sets, where features with term frequency no less than three are selected and the BOOL weight is adopted. Table 1 reports the classification accuracy of each component classifier (Base-J, Base-V, Base-N, and Base-O), and the system using all features (Unigrams), only FE, only PCA-SS, and SS-FE. We compared the following three aspects.

FE vs. Baselines. We first compare base classifiers and Unigrams. It is interesting that Base-J yields a comparative performance to Unigrams. In some tasks, Base-J is even better. With an efficient ensemble of all base classifiers, FE performs consistently better than each base classifier and the Unigrams system.

PCA-SS vs. Baselines. When comparing PCA-SS and Unigrams, we reconfirm the conclusion from Figure 1 that, with a selected subset of training samples, the PCA-SS could improve significantly.

FE, PCA-SS, and SS-FE. SS-FE is consistently better than Unigrams, FE, and PCA-SS, except for one task K → B, where SS-FE is slightly weaker than FE. We summarize the results as follows: first, in the tasks, such as D → B and K → D, where FE is more effective but the PCA-SS is less effective (denoted as $FE \succ SS$), the SS-FE improvements are generally gained by labeling adaptation. Second, in the tasks where the effects of PCA-SS are more significant than FE ($FE \prec SS$), such as E → D and E → K, the SS-FE improvements are mainly from instance adaptation. Third, in the tasks where FE and PCA-SS are both effective, such as D → E and B → K, the improvements finally gained by SS-FE are remarkable.

3.4 Why POS-Based Feature Ensemble?

Table 2 shows the average KL distance (KLD) across all tasks regarding to different POS tags and their weights learnt in FE. Generally, the KLD of different types of POS tags can be ranked as: N >>V > J > O. The KLD of N is significantly larger than the other POS tags, indicating that the change of N is the biggest across domains. The KLD of J is significantly smaller than that of N. V gives the comparable KLD. This suggests that features in J and V are more domain-independent, in comparison with N. We then observe the weights of different parts of POS tags learnt in FE. The weight of J is the largest, while the weight of N is comparatively small. This confirms our motivation for POS-based feature ensemble.

Table 2: KL distance between source and target domains in different POS tags.

POS	J	V	N	O
KLD	0.306	0.312	1.048	0.083
Weight	0.47	0.19	0.16	0.18

Table 3: KL distance between source and target domains.

Task	B-D	B-E	B-K	D-E	D-K	E-K
KLD	0.187	0.458	0.472	0.443	0.475	0.284

3.5 Why Feature Ensemble plus Sample Selection?

In this work, labeling adaptation is conducted in a feature reweighing manner by FE. Instance adaptation is embodied in the manner of sample selection, where the empirical likelihood is obtained based on the selected subset of samples. To validate this analysis, we further present the KLD of different domain pairs in Table 3. We find that the KLD of B-D and E-K are relatively low, which suggests that these domains are similar in distribution. This corresponds well with our experimental results: first, in tasks such as D → B and K → D, the improvement of SS-FE is generally gained by labeling adaptation. This is quite reasonable, because their KLD is relatively low and the demand for instance adaptation is not large. Second, for the domain pairs whose distributional change is bigger, such as E → D and E → K, the improvement of SS-FE is mainly due to instance adaptation rather than labeling adaptation.

4 Conclusions

We have presented the SS-FE approach to conduct labeling adaptation and instance adaptation together for domain adap-

tation. Experimental results showed the effectiveness of SSFE in both labeling adaptation and instance adaptation.

Acknowledgement

The work is supported by the Natural Science Foundation of China (61305090), the Jiangsu Provincial Natural Science Foundation of China (BK2012396), and the Research Fund for the Doctoral Program of Higher Education of China (20123219120025).

References

[Axelrod *et al.*, 2011] Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 355–362, 2011.

[Blitzer *et al.*, 2007] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*, volume 7, pages 440–447, 2007.

[Cambria and Hussain, 2015] Erik Cambria and Amir Hussain. *Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis*. Springer, Cham, Switzerland, 2015.

[Cambria *et al.*, 2013] Erik Cambria, Bjoern Schuller, Bing Liu, Haixun Wang, and Catherine Havasi. Knowledge-based approaches to concept-level sentiment analysis. *IEEE Intelligent Systems*, 28(2):12–14, 2013.

[Cambria *et al.*, 2014] Erik Cambria, Haixun Wang, and Bebo White. Guest editorial: Big social data analysis. *Knowledge-Based Systems*, 69:1–2, 2014.

[Cambria *et al.*, 2015] Erik Cambria, Paolo Gastaldo, Federica Bisio, and Rodolfo Zunino. An ELM-based model for affective analogical reasoning. *Neurocomputing*, 149:443–455, 2015.

[Daumé III, 2007] Hal Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 256–263, 2007.

[Jiang and Zhai, 2007] Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 264–271, 2007.

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

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

[Pan *et al.*, 2010] 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 International Conference on World Wide Web (WWW)*, pages 751–760, 2010.

[Pan *et al.*, 2011] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[Samdani and Yih, 2011] Rajhans Samdani and Wen-tau Yih. Domain adaptation with ensemble of feature groups. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1458–1464, 2011.

[Sugiyama *et al.*, 2008] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, pages 1433–1440, 2008.

[Xia and Zong, 2011] Rui Xia and Chengqing Zong. A pos-based ensemble model for cross-domain sentiment classification. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, pages 614–622, 2011.

[Xia *et al.*, 2011] Rui Xia, Chengqing Zong, and Shoushan Li. Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, 181(6):1138–1152, 2011.

[Xia *et al.*, 2013a] Rui Xia, Xuelei Hu, Jianfeng Lu, Jian Yang, and Chengqing Zong. Instance selection and instance weighting for cross-domain sentiment classification via pu learning. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2176–2182, 2013.

[Xia *et al.*, 2013b] Rui Xia, Chengqing Zong, Xuelei Hu, and Erik Cambria. Feature ensemble plus sample selection: Domain adaptation for sentiment classification. *IEEE Intelligent Systems*, 28(3):10–18, 2013.