Cross-Domain and Cross-Category Emotion Tagging for Comments of Online News

Ying Zhang[†‡]
College of Software
Nankai University
yingzhang@nankai.edu.cn

Ning Zhang[†]
Dept. of Computer Science
Purdue University
lemonustc@gmail.com

Luo Si
Dept. of Computer Science
Purdue University
lsi@cs.purdue.edu

Yanshan Lu[‡]
APEX
Twitter, Inc.
luyanshan.frank@gmail.com

Qifan Wang
Dept. of Computer Science
Purdue University
wang868@purdue.edu

Xiaojie Yuan
College of Software
Nankai University
yuanxj@nankai.edu.cn

ABSTRACT

In many online news services, users often write comments towards news in subjective emotions such as sadness, happiness or anger. Knowing such emotions can help understand the preferences and perspectives of individual users, and therefore may facilitate online publishers to provide more relevant services to users. Although building emotion classifiers is a practical task, it highly depends on sufficient training data that is not easy to be collected directly and the manually labeling work of comments can be quite labor intensive. Also, online news has different domains, which makes the problem even harder as different word distributions of the domains require different classifiers with corresponding distinct training data.

This paper addresses the task of emotion tagging for comments of cross-domain online news. The cross-domain task is formulated as a transfer learning problem which utilizes a small amount of labeled data from a target news domain and abundant labeled data from a different source domain. This paper proposes a novel framework to transfer knowledge across different news domains. More specifically, different approaches have been proposed when the two domains share the same set of emotion categories or use different categories. An extensive set of experimental results on four datasets from popular online news services demonstrates the effectiveness of our proposed models in cross-domain emotion tagging for comments of online news in both the scenarios of sharing the same emotion categories or having different categories in the source and target domains.

[†]The authors contributed equally to this work.

[‡]Part of this work was performed when the authors visited/studied at Purdue University.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609587.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval - Information Search and Retrieval

Keywords

Sentiment Tagging; Online News; Comments; Transfer Learning; Cross-Domain; Cross-Category

1. INTRODUCTION

With the explosion of social media over the past decade, more and more user-generated data is available on the Web for expressing users' opinions and emotions. Among various types of social media, online news is an important form that attracts billions of users to read, respond, and actively interact with each other by making comments. Opinions and comments of individual users often have huge impacts on other users and the community. Users often express subjective emotions such as sadness, surprise and anger in comments. Grasping such emotions can help understand the perspectives and preferences of individual users, and thus may facilitate online publishers to provide more personalized services or statistically study readers' attitudes toward news events. Therefore, to better make use of user comments for quantitative analysis, a research problem of automatic emotion tagging for comments of online news arises.

Emotion tagging for online news comments is an application of the research area of opinion mining and sentiment analysis, which has attracted much attention in information retrieval and natural language processing communities (e.g., [24]). In particular, the tagging problem is formulated as a sentiment classification problem, which focuses on detecting the polarity (e.g., positive or negative) or multiple emotion categories (e.g., happy, sad, angry, etc.) from user-generated contents including user reviews of products or services, posts on blogs or social networks, and comments in forums or online news services.

Traditional supervised learning methods have been applied to emotion tagging for comments of online news (e.g., [34]). The performances of these methods heavily rely on the availability of a relative large amount of manually tagged comments. The labeling work is often labor intensive for obtaining sufficient training data. Moreover, online news has many domains such as politics, entertainment or sports, and if we directly apply the classifiers trained from one domain to comments from another domain, it usually leads to poor classification performances (e.g., [5][22]). The reason is that comments in different news domains talk about different sets of topics in different styles, which results in different term distributions.

This paper focuses on the task of cross-domain emotion tagging, which utilizes the model trained on one source news domain to help build the model for another target news domain. More specifically, there are two different scenarios in the task of cross-domain emotion tagging for online news comments. The first scenario is that the two domains share the same set of emotion categories (e.g., both are emotion polarities). The second one is that the two domains use different emotion categories which lead to different label spaces. For example, comments in the source domain are tagged with binary sentiment polarity whereas in the target domain we want to tag comments with multiple emotion categories such as happiness, sadness, and anger. Taking the category differences into consideration, the problem turns to be cross-domain and cross-category emotion tagging.

This paper proposes novel transfer learning approaches for cross-domain and cross-category emotion tagging of online news comments. When abundant labeled data in the source domain are available, a relatively small amount of labeled data in the target domain can be sufficient for training an accurate classifier in the target domain by transferring useful knowledge from the source domain. In particular, when two domains share the same set of emotion categories, the joint probabilities of text features and emotion categories are used to model the relationship between data in the source and target domains. When the two domains utilize different sets of categories, we propose two types of models which deal with the category differences either in a probabilistic way or in an explicit way: the probabilistic model infers the category correlations purely from data and the explicit models borrow human prior knowledge and fix the matchings of categories.

An extensive set of experiments is conducted with four datasets from popular online news services in two groups. Our proposed models significantly outperform alternatives in the task of emotion tagging for comments in both cases when the source domain and target domain share the same emotion categories or use different categories.

To the best of our knowledge, this is the first piece of research that focuses on modeling cross-domain and cross-category emotion tagging for comments of online news. Our proposed models can also be generalized and applied on other cross-domain and cross-category applications.

The rest of the paper is organized as follows. Section 2 reviews some related work. Section 3 proposes our approaches for the two scenarios when the source domain and the target domain share the same emotion categories or use different categories. Datasets and experimental results are discussed in Section 4. The last section provides conclusions and points out possible future work.

2. RELATED WORK

Sentiment analysis has become an important subfield of information management. Previous work mostly focuses on product reviews [21], blogs [7][20][30] and news corpora [2][8][19]. For a general survey, please refer to [24]. This paper focuses on emotion tagging for comments of online news.

Many machine learning techniques have been applied on sentiment classification, such as unsupervised learning techniques (e.g., [28]), supervised learning techniques (e.g., [25]) and semi-supervised learning techniques (e.g., [26]). For comments of online news, Zhang *et al.* [34] propose a classification approach that exploits information from heterogeneous information sources to predict emotions. The research work in [15] tries to predict upcoming comments' polarity before comments are posted. Many existing works on sentiment analysis make a strong assumption that the labeled training data and unlabeled testing data share the same features and category spaces. Thus to analyze comments from a new domain, the training data need to be recollected and models need to be rebuilt.

But the tasks and data distributions in those domains are not completely different, transfer learning techniques [23][29] can reuse data from an existing domain and thus require less labeled data from the new domain. Blitzer *et al.* [3][4] propose the structural correspondence learning (SCL) algorithm to exploit domain adaptation techniques for sentiment classification. SCL is motivated by a multi-task learning method using alternating structural optimization (ASO) [1]. More recently, Li *et al.* [18] propose to transfer common lexical knowledge across domains via matrix factorization techniques. Zhang *et al.* [32] propose research on knowledge transfer in sentiment analysis with auxiliary data from related sources. Unlike SCL algorithm which uses source domain data to find some important pivot features, it models the underlying distribution differences explicitly. The work of Zhang *et al.* is designed for binary categories and can handle neither multiple categories nor the scenario when two domains use different category sets.

3. CROSS-DOMAIN AND CROSS-CATEGORY EMOTION TAGGING

This section proposes a novel framework of cross-domain and cross-category emotion tagging for comments of online news by modeling the data differences between the source and target domains. First we provide a formal definition of the research problem, and then propose approaches for the two scenarios when the source and target domains share the same set of emotion categories or use different categories.

3.1 Problem Definition

Given a specific news domain D, where an emotion category set is defined as $\mathcal{E} = \{e_i\}(i = 1, \dots, k)$, the comment emotion tagging problem is to tag unlabeled comments with corresponding emotions in $\mathcal{E}$ utilizing a set of labeled comments $\mathcal{D}$. A comment is described as a feature vector $\mathbf{x}$ and the label is defined as $\mathbf{y} = \{y_l\}(l = 1, \dots, K)$, y_l equals to 1 if comment $\mathbf{x}$ is tagged with emotion e_l or 0 otherwise, $\sum_{l=1}^{K} y_l = 1$. Thus the labeled comments set $\mathcal{D}$ can be denoted as $\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_m, \mathbf{y}_m)\}$, where m is the number of labeled comments.

In our problem setting of cross-domain and cross-category emotion tagging, two news domains D_S and D_T are specified, where D_S is the source domain and D_T is the target domain. Two emotion category sets $\mathcal{E}_S = \{e_{S_l}\}(l = 1, \dots, K_S)$ and $\mathcal{E}_T = \{e_{T_k}\}(k = 1, \dots, K_T)$ are defined on D_S and D_T respectively. Notice that here $\mathcal{E}_S$ and $\mathcal{E}_T$ are not necessarily to be the same and it is the reason why our setting is described by "cross-category". We have a set of labeled data from the source domain as $\mathcal{D}_S = \{(\mathbf{x}_{S_1}, \mathbf{y}_{S_1}), (\mathbf{x}_{S_2}, \mathbf{y}_{S_2}), \ldots, (\mathbf{x}_{S_m}, \mathbf{y}_{S_m})\}$, where m is the size of source domain data. In addition, some labeled data from the target (the new) domain as $\mathcal{D}_T = \{(\mathbf{x}_{T_1}, \mathbf{y}_{T_1}), (\mathbf{x}_{T_2}, \mathbf{y}_{T_2}), \ldots, (\mathbf{x}_{T_n}, \mathbf{y}_{T_n})\}$ is also available and n is the size of labeled target domain data. $\mathbf{x}_{S_i} \in \mathcal{X}_S$ and $\mathbf{x}_{T_i} \in \mathcal{X}_T$ are drawn from the same feature space but under different distributions. In general, $0 \leq n \ll m$ as there are abundant labeled data in the source domain but limited labeled data in the target domain. The task of cross-domain and cross-category emotion tagging is to learn an emotion classifier to predict the emotion tags of unlabeled comments in the target domain D_T based on the labeled data of both $\mathcal{D}_S$ and $\mathcal{D}_T$.

3.2 Emotion Tagging with the Same Categories

This section focuses on the problem of cross-domain emotion tagging with identical category sets in D_S and D_T. Namely, $\mathcal{E}_S$ is exactly the same as $\mathcal{E}_T$ and we rewrite them as $\mathcal{E}_S = \mathcal{E}_T = \{e_k\}(k = 1, \ldots, K)$.

3.2.1 Learning Objective

Many probabilistic classification techniques in the literature generally fall into two categories: generative models and discriminative models. Discriminative models have attractive theoretical properties [16] and their effectiveness has been demonstrated in the field of information retrieval for applications such as text classification [14][31] and learning to rank [13]. In most cases, discriminative models provide better accuracy than generative models.

This paper utilizes multinomial logistic regression, a discriminative model, to classify users' comments into different emotion categories. Formally, in the target domain D_T and given the i^{th} comment, the conditional probability that the comment should be tagged with a predefined emotion category e_k is expressed in terms of a soft-max function as the following normalized probabilistic value,

$$\psi_{T_{ik}} = P(e_k|x_{T_i}) = \frac{\exp(\omega_k^T x_{T_i})}{\sum_{r=1}^{K} \exp(\omega_r^T x_{T_i})} \tag{1}$$

Here $\mathbf{x}_{T_i}$ represents the feature vector of the i^{th} comment, ω_k denotes the combination weight parameters of emotion category e_k. The larger the value $\omega_k^T x_{T_i}$ is, the more probable the comment shows particular emotion e_k.

The error of the emotion tagging approach with multinomial logistic regression is measured with the negative log-likelihood loss function. In particular, the loss of a particular i^{th} instance in the training data of the target domain, denoted by ξ_{T_i}, can be expressed as follows,

$$\xi_{T_i} = -\log \prod_{k=1}^{K} \psi_{T_{ik}}^{y_{T_{ik}}} = -\sum_{k=1}^{K} y_{T_{ik}} \log \psi_{T_{ik}} \tag{2}$$

Since the training data from the target domain are insufficient for making accurate prediction, it is necessary to utilize the abundant training data from the source domain to improve the classification accuracy in the target domain. Although the feature spaces are the same (i.e., a vector space of keywords in the vocabulary), the data distributions of the two domains are different as the comments in the source domain and the comments in the target domain focus on different types of topics.

To incorporate the auxiliary training data from the source domain into the loss function of the target domain, the differences of training data distributions between different domains are modeled. We propose to adjust the weight of labeled data from the source domain to reflect their corresponding importance in the target domain. In particular, a unified objective loss function is defined as follows, taking data from both the source domain and the target domain into account:

$$\min_{\omega} \frac{\lambda_1}{n} \sum_{i=1}^{n} \xi_{T_i} + \frac{\lambda_2}{m} \sum_{i=1}^{m} \beta_i \xi_{S_i} + R(\omega) \tag{3}$$

Where ξ_{S_i} is the loss value of each instance from the source domain, which is defined with respect to the weight parameters ω_k as follows (i.e., in a similar manner to the loss function in the target domain),

$$\xi_{S_i} = -\sum_{k=1}^{K} y_{S_{ik}} \log \frac{\exp(\omega_k^T x_{S_i})}{\sum_{r=1}^{K} \exp(\omega_r^T x_{S_i})} \tag{4}$$

In equation (3), β_i models the data distribution differences between the source domain and the target domain, which enables the source domain instances to be naturally incorporated into the objective function through appropriate reweighting. The estimating method of β_i is introduced in the next subsection. $R(\omega)$ is the regularization penalty to prevent overfitting. In particular, we use the L2 (i.e., ridge) regularization method [12]. λ_1 and λ_2 are two trade-off parameters that explore the relative importance of classification results in the source domain and the target domain.

3.2.2 Distribution Differences between Domains

As we have discussed, domain adaption can be achieved through reweighting instances from the source domain by modeling the probability differences between the distributions of domains. The distribution differences between the source domain and the target domain lie not only in the text features (feature spaces) of comments but also in the emotion categories (label spaces). So we use the joint probabilities of the text features and the emotion categories to model the relationship between data in the source and target domains. In particular, β_i refers to $\frac{Pr_T(\mathbf{x}_{S_i}, \mathbf{y}_{S_i})}{Pr_S(\mathbf{x}_{S_i}, \mathbf{y}_{S_i})}$, which represents the ratio between the joint data distributions in the target domain and that in the source domain for the source domain instance $\mathbf{x}_{S_i}$. The intuition is that the ratio of the joint probabilities can well capture how a data instance from the source domain should be reweighted to reflect its importance in the target domain.

To estimate β_i, we adopt an approach based on kernel density estimation (many other methods such as kernel mean matching and Gaussian mixture model can also be utilized). In particular, β_i is formulated as follows,

$$\beta_i = \frac{Pr_T(\mathbf{x}_{S_i}, \mathbf{y}_{S_i})}{Pr_S(\mathbf{x}_{S_i}, \mathbf{y}_{S_i})} = \frac{Pr_T(\mathbf{x}_{S_i}|\mathbf{y}_{S_i}) \times Pr_T(\mathbf{y}_{S_i})}{Pr_S(\mathbf{x}_{S_i}|\mathbf{y}_{S_i}) \times Pr_S(\mathbf{y}_{S_i})} \tag{5}$$

It is clear that $\frac{Pr_T(\mathbf{y}_{S_i})}{Pr_S(\mathbf{y}_{S_i})}$ represents the ratio of emotion category probabilities between the target domain and the source domain, which can be estimated from the labeled data on both domains based on the relative frequencies of each emotion category. $\frac{Pr_T(\mathbf{x}_{S_i}|\mathbf{y}_{S_i})}{Pr_S(\mathbf{x}_{S_i}|\mathbf{y}_{S_i})}$ can be estimated by kernel density estimation with the Gaussian kernel as follows,

$$\frac{\frac{1}{|\sum_{j=1}^{n} I_{T_{ij}}|}\sum_{j=1}^{n} I_{T_{ij}} \exp(-\frac{\|\mathbf{x}_{S_i}-\mathbf{x}_{T_j}\|}{\sigma^2})}{\frac{1}{|\sum_{j=1}^{m} I_{S_{ij}}-1|}(\sum_{j=1}^{m} I_{S_{ij}} \exp(-\frac{\|\mathbf{x}_{S_i}-\mathbf{x}_{S_j}\|}{\sigma^2})-1)} \quad (6)$$

Where $I_{T_{ij}}$ is an indicator function, which equals to 1 if $\mathbf{y}_{S_i} = \mathbf{y}_{T_j}$ or 0 otherwise. $I_{S_{ij}}$ is a similar indicator function, which equals to 1 if $\mathbf{y}_{S_i} = \mathbf{y}_{S_j}$ or 0 otherwise. σ is the bandwidth parameter for the Gaussian kernel. The -1 factor in the denominator removes the effect of the instance itself (i.e., $(\mathbf{x}_{S_i}, \mathbf{y}_{S_i})$) in the source domain for modeling the probability ratio.

It can be seen that if a source domain instance is close enough to target domain instances, its importance will be high. This is consistent with our expectation, which means the corresponding instance is more representative in the target domain. With this approach, the data distribution of the training data in the source domain can be adjusted to fit the data distribution in the target domain.

The objective loss function expressed by Equation (3) forms a smooth convex optimization problem, and it can be optimized by any gradient descent method. In particular, we use the Quasi-Newton method [6] that enjoys fast convergence rate with limited storage requirement.

3.3 Emotion Tagging on Different Categories

The model introduced in the last section works in the scenario when the source domain and the target domain share the same set of emotion categories. On the other side, it is also necessary to consider domains associated with different emotion categories. For example, the source domain may contain a larger amount of binary sentiment polarity data (e.g., positive and negative) that can be more easily labeled, whereas in the target domain we want to tag multiple emotion categories (e.g., happy, surprised, sad and angry) and only have a limited amount of labeled training data. For building cross-domain emotion tagging solutions in this scenario, it is necessary to model the relationship between different sets of emotion categories in the source domain and the target domain to deal with the differences of label spaces.

Section 3.3.1 introduces a probabilistic model which infers the category correlations from data and Section 3.3.2 presents explicit models in which the category matchings are explicitly fixed by incorporating human prior knowledge.

3.3.1 A two-level probabilistic model

Emotion categories are often highly correlated. For example, one set is a coarse-grained polarity set with two categories of positive and negative, and the other set is a fine-grained emotion set containing happy, amused, angry, and sad. Even though they are different, we can easily find some correlations: if a comment shows happy or amused, it tends to be tagged as positive, and if it is labeled as angry or sad, it tends to have negative sentiment.

This nature of emotion categories inspires us to develop a model that can utilize the correlation between emotion categories to transfer knowledge: first classify the comments from the target domain into the set of emotion categories of the source domain, and then use those class probabilities generated in the first step as features to infer the emotion category in the target domain. This is a two-level probabilistic model, in which we make use of abundant source domain data in the first level, and infer emotion category correlations in the second level.

The first step is similar to the problem addressed in Section 3.2, thus it is possible to utilize a similar model for transferring knowledge from the source domain within the first step. More specifically, given the i^{th} comment in the target domain, the conditional probability that the comment should be associated with a predefined emotion e_{S_l} $(l = 1, \ldots, K_S)$ within the set of emotion categories in the source domain can be calculated as normalized probability values with a soft-max function as follows,

$$\delta_{T_{il}} = P(e_{S_l}|x_{T_i}) = \frac{\exp(\omega_l^T x_{T_i})}{\sum_{r=1}^{K_S} \exp(\omega_r^T x_{T_i})} \quad (7)$$

Here $\mathbf{x}_{T_i}$ represents the feature vector of the i^{th} data instance (i.e., user comment) in the target domain, and ω_l denotes the weight parameters of a particular emotion e_{S_l}.

In the second step, we take the outputs of the classifier above as features, namely, δ_{T_i} is composed by $\delta_{T_{il}}$ $(l = 1, \ldots, K_S)$, and use multinomial logistic regression model to transfer knowledge between different emotion categories in different domains. The negative loglikelihood function can be defined as follows:

$$\xi'_{T_i} = -\sum_{k=1}^{K_T} y_{T_{ik}} \log \frac{\exp(\nu_k^T \delta_{T_i})}{\sum_{r=1}^{K_T} \exp(\nu_r^T \delta_{T_i})} \quad (8)$$

Here ν_k represents the weight parameters of a particular emotion e_{T_k}.

By incorporating the auxiliary training data from the source domain in the first level, the objective loss function can be defined as follows:

$$\min_{\omega,\nu} \frac{\lambda_1}{n}\sum_{i=1}^{n} \xi'_{T_i} + \frac{\lambda_2}{m}\sum_{i=1}^{m} \gamma_i \xi_{S_i} + \lambda_3 R(\nu) + R(\omega) \quad (9)$$

where ξ_{S_i} is the loss value of each instance in the training data from the source domain, defined in the same way of Equation (7). γ_i models the data distribution differences between the source domain and the target domain for reweighting the source domain instances. $R(\nu)$ and $R(\omega)$ are two regularization penalties for parameters ν and ω respectively, in order to prevent overfitting. We use the L2 (i.e., ridge) regularization method. λ_1, λ_2 and λ_3 are three trade-off parameters. Note that if either the number of emotion categories in the target domain or that in the source domain is two, we can simply replace the soft-max function by binary sigmoid function accordingly.

Due to the differences of emotion categories between domains, we are no longer able to model the joint data distribution in different domains. Instead, the marginal probabilities of text features of comments are used. Specifically, γ_i is defined with Gaussian kernel density estimator as follows, in a similar way as described in Section 3.2.2.

$$\frac{Pr_T(\mathbf{x}_{S_i})}{Pr_S(\mathbf{x}_{S_i})} \propto \frac{\frac{1}{n}\sum_{j=1}^{n} \exp(-\frac{\|\mathbf{x}_{S_i}-\mathbf{x}_{T_j}\|}{\sigma^2})}{\frac{1}{m-1}(\sum_{j=1}^{m} \exp(-\frac{\|\mathbf{x}_{S_i}-\mathbf{x}_{S_j}\|}{\sigma^2})-1)} \quad (10)$$

There are two sets of parameters to be estimated in the joint optimization problem, ω and ν, and we employ an alternative optimization method for the task. The value of ν can be fixed, and then the problem is converted to a convex optimization problem with respect to ω, which can be solved by any gradient descent method. The Quasi-Newton method [6] is used for this purpose. After that, the value of ω is fixed and the value of ν can be optimized in a similar manner. The above alternative process can be conducted iteratively until the convergence for obtaining optimal values of $\omega^\star$ and $\nu^\star$. The learned model can then be used for emotion tagging of any new comment in the target domain.

3.3.2 Explicit models of category correlation

In the above two-level probabilistic model, we first tag comments in the source domain label space, in order to make use of abundant source domain data, and then infer the target domain labels from those source domain class probabilities. However, it may not work well when transferring knowledge from a coarse-grained source domain to a fine-grained target domain, because in the second level it can be too difficult to infer fine-grained categories only from coarse-grained category probabilities.

On the other side, the valuable human prior knowledge about the correlation of category sets (e.g. which emotion categories are positive and which are negative) can be utilized in a more explicit manner. We propose the explicit models of category correlation which overcomes the deficiency of the probabilistic model.

First we formalize the explicit category correlation named matching correlation. Given a coarse-grained emotion category set $\mathcal{E}_c = \{e_i\}(i = 1, \ldots, K_c)$ and a fine-grained emotion category set $\mathcal{E}_f = \{e_i\}(i = 1, \ldots, K_f$ and $K_f > K_c)$, if we can define a matching function $M : \mathcal{E}_f \to \mathcal{E}_c$ based on reasonable relationship of emotion categories, we say $\mathcal{E}_c$ and $\mathcal{E}_f$ have the matching correlation. For example, in the emotion category set with "happy", "amused", "angry", and "sad" and the sentiment polarity set, happy and amused can be matched to positive and angry and sad can be matched to negative, so these two sets have matching correlation.

Sub-Scenario 1: when the source domain is fine-grained and the target domain is coarse-grained and the emotion category sets have matching correlation, we propose a one-level explicit model as follows: first match the fine-grained labels of the source domain data to corresponding labels in the coarse-grained target domain category set according to matching function M and get a matched coarse-grained source domain dataset $\mathcal{D}'_S$, then we directly apply the model introduced in Section 3.2 to transfer knowledge from $\mathcal{D}'_S$ and do emotion tagging in the target domain label space with the target dataset $\mathcal{D}_T$. In this model we abandon some unnecessary fine-grained label information of the source domain data and it helps reduce modeling noise.

Sub-Scenario 2: when the source domain is coarse–grained and the target domain is fine-grained and the emotion category sets have matching correlation, we propose a two-level explicit model. In the first level, we match the fine-grained labels of target domain data to corresponding labels in the coarse-grained source domain category set according to function M and get a matched coarse-grained target domain dataset $\mathcal{D}'_T$, then we apply the model introduced in Section 3.2 to transfer knowledge from $\mathcal{D}_S$ and do emotion tagging in the course-grained source domain label space with the matched target dataset $\mathcal{D}'_T$. With this first step, we transfer knowledge from $\mathcal{D}_S$ and build a model which can more accurately tag comments in the target domain with coarse-grained labels as intermediate results.

In the second level, K_c sub-classifiers are built for each category e_i in the coarse-grained $\mathcal{E}_S$ to further classify the fine-grained categories which are matched to the same coarse-grained category e_i. More specifically, $\mathcal{D}_T$ are divided into K_c sub-sets according to the labels in the coarse-grained $\mathcal{D}'_T$, which means comments in each sub-set have the same coarse-grained tag in $\mathcal{D}'_T$. Then sub-classifiers are built for each sub-set to classify corresponding fine-grained emotions. Multinomial logistic regression models with L2 regularization method are applied for building those sub-classifiers.

In this two-level explicit model, we actually regard the tagging problem as a two-step procedure: first classify data into coarse-grained categories (e.g., polarities), then classify fine-grained categories within each coarse-grained category. The explicit model improves the tagging accuracy through transferring source domain knowledge in the first step.

We name the above two models with "explicit" because the category correlations are used explicitly as prior knowledge instead of learned purely by data. By explicitly matching the categories, the explicit models directly project data from one domain into the label space of another domain and hence reduce noise and avoid unnecessary uncertainty.

4. EXPERIMENTAL EVALUATION

In this section, we first introduces the experimental datasets and analyzes the domain differences between datasets. Then we reports an extensive set of experimental results of our proposed approaches and baseline algorithms in both scenarios of using the same emotion categories and different categories in the source and target domains. Analysis and discussions are presented based on the results.

4.1 Experimental Setups

4.1.1 Datasets

Two groups of datasets are used conducting the experiments and each group contains two datasets from different news domains. More specifically, in every run of experiments, one dataset in a group acts as the target domain dataset and the other one in the group acts as the source domain dataset.

The first group of datasets (first used in [34]) is in Chinese and comes from two online news services as Sina News and QQ News, which are among the largest news portals in China. In particular, top 20 most popular comments of most-viewed news articles are collected within six months of 2011 from the Society channel of Sina News and the Entertainment channel of QQ News. These two datasets are referred as the Sina Society dataset and the QQ Entertainment dataset respectively.

The second group of datasets is in English and is collected (by Bruno Jakic [15]) from the famous social news portal Reddit. More than $20K$ news posts and corresponding comments from 8 domains are collected during the time period from January 2011 to March 2011. This paper choose 4 domains which are Politics, WorldNews, Science, and Technology. Moreover, Politics and WorldNews comments are merged together and referred as Reddit Poli&WorldNews dataset and Science and Technology comments are merged together and referred as Reddit Sci&Tech dataset.

Emotion labels in all of the four datasets are manually annotated. In Sina News and QQ News, even though readers can tag articles with built-in emotion categories, the tagging systems are independent from the commenting systems so a tag cannot be paired with a specific comment. Meanwhile, users provide much fewer emotion tags than comments, probably because that most users feel comment is a better way for expressing their feelings. Thus we cannot utilize users' tags as labels and instead, we just borrow the built-in emotion categories as predefined fine-grained emotion categories in the annotating task for Sina Society dataset and QQ Entertainment dataset (note that the two category sets are not the same). For the two Reddit datasets, no built-in emotion category is provided and we pick six emotions out from the 6 basic emotion categories [10] and some other common emotion categories as fine-grained emotion categories. For all of the four datasets, positive/negative are used as coarse-grained categories.

Due to the substantial laboring efforts, each dataset is annotated by only two annotators as one takes charge of the coarse-grained polarity annotating and the other one conducts the fine-grained emotion annotating. Neutral comments are excluded since we focus on emotion classification instead of subjectivity detection. The statistics of annotated datasets are as shown in Table 1 and Table 2.

Table 1: The statistics of labeled comments on the 8 emotion categories of Chinese datasets.

Sina Society		QQ Entertainment	
Category	Number	Category	Number
Touched	899	Touched	139
Sympathetic	612	Sympathetic	643
Angry	1743	Angry	1641
Amused	409	Amused	564
Sad	656	Sad	358
Surprised	195	Surprised	86
Fervent	322	Happy	1626
Bored	338	Anxious	374
Positive	2028	Positive	2494
Negative	3146	Negative	2937
Total	5174	Total	5431

Table 2: The statistics of labeled comments on the 8 emotion categories of Reddit datasets.

Poli&WorldNews		Sci&Tech	
Category	Number	Category	Number
Sympathetic	300	Sympathetic	125
Angry	619	Angry	556
Disgust	687	Disgust	704
Surprised	699	Surprised	461
Happy	628	Happy	1000
Sad	178	Sad	287
Positive	1668	Positive	1688
Negative	1631	Negative	1681
Total	3299	Total	3369

To test the annotating quality, 100 comments are randomly sampled from each dataset and a reviewer (not the annotator) annotated them blindly from the original labels. The number of consistent labels are listed in Table 3.

4.1.2 Domain difference

For different datasets in the same group, distributions in either the emotion label spaces or the word feature space are distinct since they are from different domains. Table 1 and Table 2 show the distribution differences in the label spaces.

Table 3: Number of consistent labels in 100 samples

Dataset	coarse-grained	fine-grained
Sina Society	99	91
QQ Entertainment	98	94
Reddit Poli&WorldNews	100	95
Reddit Sci&Tech	96	89

The two Chinese datasets have different category sets while the two Reddit datasets share the same one. It leads us to choose the 6 overlapped categories for the experiment of same category emotion tagging on Chinese datasets.

For feature spaces, emotion terms have been commonly used as features in the task of textual emotion recognition [27], since they are more likely to convey the emotions. This is consistent with our observation that more accurate results of textual emotion recognition can be obtained using only textual emotion features than using all bag-of-word features. Therefore, we utilize the occurrences of emotion terms in the content of comments as features. For the Chinese datasets, we use a Chinese word segmentation software ICTCLAS [33] to segment the comments into terms and extract the emotion terms by two lexical resources (i.e., NTU Sentiment Dictionary [17], Hownet [9]). For the Reddit datasets which are in English, we use the Python NLTK toolkit to preprocess the comments and extract the emotion terms based on SentiWordNet [11]. As an example from Redddit Poli&Worldnews dataset, the feature terms of a comment "Reactor pool 4 was empty and the disaster is worse than Chernobyl" are "pool", "empty", "disast", and "wors".

Although the two Chinese datasets share the same Chinese emotion term set as feature space (i.e., vocabulary) and the two Reddit datasets share the same English emotion term set as feature space, their feature distributions vary from each other. Figure 1 shows the differences of feature distributions between the two Chinese datasets. The X axis represents the term features which are ordered according to their term frequencies in QQ Entertainment dataset. Then the ln(term frequencies) values of both of the two Chinese datasets are plotted. The comparison of the feature distributions of the Reddit datasets is similar.

4.1.3 Evaluation Metrics

We adopt Accuracy($Accu@m$) as the measurement. Given a comment c_i, its labeled emotion $\bar{e}_i$ and predicted emotion set $\mathcal{E}_i@m$ including top m ranked emotions, define

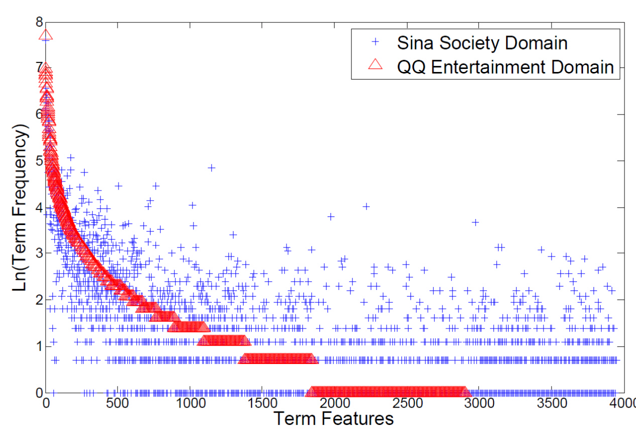

Figure 1: Comparison between feature distributions of Chinese datasets

$$accu_i@m = \begin{cases} 1, & \bar{e}_i \in \mathcal{E}_i@m \\ 0, & \bar{e}_i \notin \mathcal{E}_i@m \end{cases}$$

$Accu@m$ for the entire collection $\mathcal{C} = \{c_i\}(i = 1, \ldots, N)$ is defined as follows: $Accu@m = \frac{1}{N}\sum_{i=1}^{N} accu_i@m$.

$Accu@2, 3$ can be calculated only under the settings where the target domain has more than 2 emotion categories.

4.2 Experiments with Same Emotion Categories

Experiments in this section investigate the effectiveness of our proposed approach and baseline methods for the scenario when source domain and target domain share the same set of emotion categories.

The following methods are compared:

- **Cross-Domain Emotion Tagging with Joint Probabilities (CDET_J).** The approach proposed in Section 3.2, which reweight source domain data based on joint probabilities of text features of comments and emotion categories.
- **Structural Correspondence Learning (SCL).** A state-of-art method for transfer learning proposed by Blitzer *et al.* [3]. It is implemented with logistic regression and the pivot features are chosen as the shared top 200 frequently occurred terms.
- **Emotion Tagging by Logistic Regression (ETLR).** An emotion tagging method for comments of online news proposed in [34]. It is based on multinomial logistic regression model with L2 regularization. This method does not involve transfer learning techniques.

For all the methods, the trade-off parameters are set by five fold cross-validations. The average experimental results of 20 independent runs are reported.

The performance of all methods is evaluated on the two groups of datasets and each group has two settings by choosing either dataset as the target domain dataset. For the Chinese datasets, we use all data labeled in the overlapped 6 emotion categories in the source domain and randomly selected 1/64, 1/32, 1/16, 1/8 and 1/4 of data labeled in the target domain as the training data, and the remaining data in the target domain are used for testing. Figure 2 and Figure 3 show the Accu@1 results on both two groups of datasets with different ratios of training data in the target domain. In particular, detailed Accu@1,2,3 results with 1/16 training samples on Sina Society dataset as the target domain dataset and QQ Entertainment dataset as the source domain dataset are further reported in Table 4.

The results from Figure 2 and Figure 3 show that the performance of all three methods improves with more training data in the target domain, which is as expected. It can be seen from these results that under most settings the two methods utilizing data from both the source and target domains beat ETLR which only uses data from the target domain especially when training data from the target domain is not sufficient. This fact clearly demonstrates the advantage of transferring knowledge from the source domain. The proposed approach CDET_J outperforms SCL in most cases. This is consistent with our expectation that modeling the joint distribution helps capture useful information in data across domains. In most cases, SCL performs only slightly better than ETLR. This might be because SCL does not explicitly model the domain distribution differences and it highly depends on auxiliary tasks.

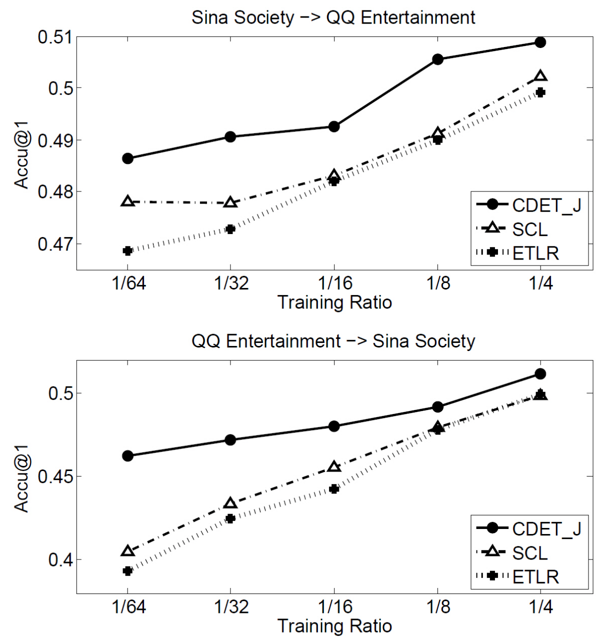

Figure 2: The Accu@1 results with different training ratios in the target domain with the same emotion categories on Chinese datasets.

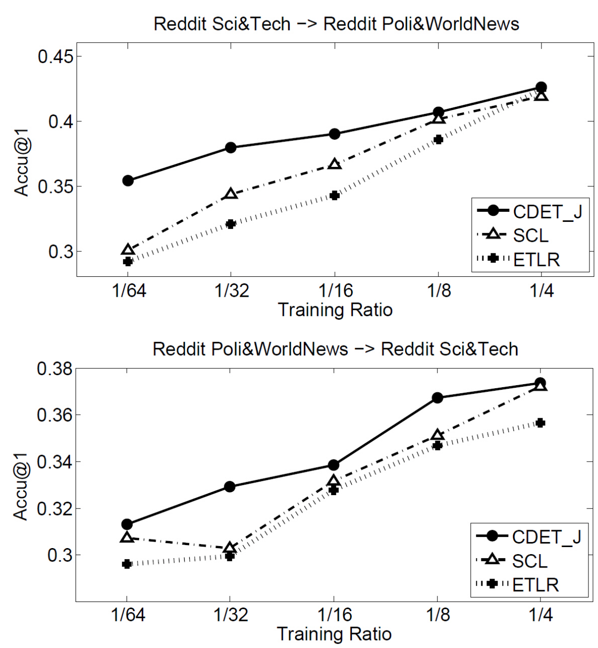

Figure 3: The Accu@1 results with different training ratios in the target domain with the same emotion categories on Reddit datasets.

Table 4: The Accu@1,2,3 results of 1/16 training ratio, QQ Entertainment is the target domain with the same emotion categories as in the source domain.

Methods	Accu@1	Accu@2	Accu@3
CDET_J	0.4925	0.6831	0.7775
SCL	0.4832	0.6829	0.7621
ETLR	0.4821	0.6044	0.7564

Significance tests using t-distribution are conducted for the Accu@1 results regarding result from each run as a sample. Benefited from sufficient independent runs, such tests provide good discriminative power. On Chinese datasets, CDET_J outperforms SCL and ETLR with 0.95 confidence level under all training ratios. On Reddit datasets, similar comparison results hold with training ratios lower or equal to 1/16. (Detailed p-values are not presented due to the limitation on number of pages.)

ETLR models trained with only source domain data are also evaluated on target datasets. It performs much worse than CDET_J and also worse than ETLR trained with sufficient (e.g. more than 1/32) target domain data, we just claim it here instead of presenting detailed results.

4.3 Experiments with Different Emotion Categories

Experiments in this section investigate the effectiveness of our proposed new methods for the scenario when the source and the target domains use different sets of emotion categories. The following methods are compared:

- **Explicit Models of Cross-Domain and Cross-Category Emotion Tagging (E_CDCCET).** The approach proposed in Section 3.3.2. By explicitly utilizing human prior knowledge on category relations and then applying CDET_J, this approach overcomes the deficiency of P_CDCCET.
- **Probabilistic Model of Cross-Domain and Cross-Category Emotion Tagging (P_CDCCET).** The approach proposed in Section 3.3.1. It models the relationship between different sets of emotion categories in different domains in a probabilistic way.
- **Emotion Tagging by Logistic Regression (ETLR).** Exactly the same as what in the previous section. Model is learned with only the data from the target domain.

For all models, their trade-off parameters are set by 5 fold cross-validations. The average experimental results of 20 independent runs are reported. Notice that the SCL method is not compared as baseline here, as it cannot fulfill the requirement of transfer learning between different label spaces.

In the following experiments, all data from the source domain are available for training, and different amounts of data are randomly sampled from the target domain for training while the remaining are for testing.

4.3.1 *Fine-grained source domain and coarse-grained target domain*

We first present the experiment results when the source domain is fine-grained and the target domain has binary polarities. For the Chinese datasets, it is no longer required to use overlapped categories so each of them contains 8 emotion categories. Figure 4 and figure 5 show the experimental results with different sizes of target domain training data.

From the results, it can be observed that both of the two proposed models outperform ETLR under most of the settings. This fact again clearly shows the effectiveness of transferring knowledge, even when it is from a fine-grained source domain to a coarse-grained target domain. Meanwhile, the explicit model E_CDCCET outperforms the probabilistic model P_CDCCET, which proves that incorporating prior knowledge on emotion category correlation helps improve classification accuracy.

Statistical significance tests have also been conducted for the Accu@1 results. On all domain settings, E_CDCCET outperforms ETLR with 0.95 confidence level under all training ratios and P_CDCCET outperforms ETLR with 0.95 confidence level under training ratios lower or equal to 1/32. Meanwhile, E_CDCCET outperforms P_CDCCET with 0.95 confidence level in most cases.

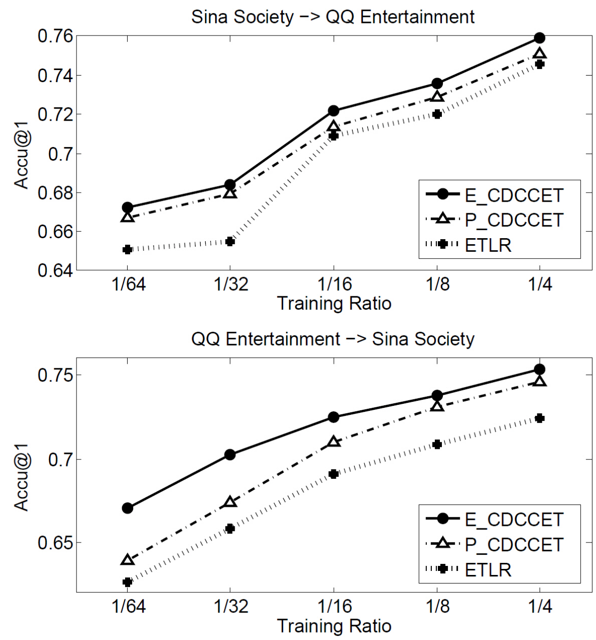

Figure 4: The Accu@1 results with fine-grained source domain and coarse-grained target domain on Chinese datasets.

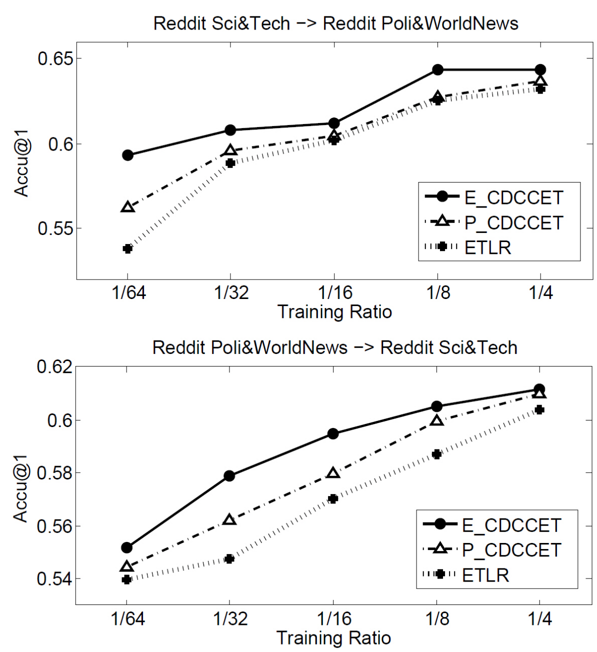

Figure 5: The Accu@1 results with fine-grained source domain and coarse-grained target domain on Reddit datasets.

4.3.2 *Coarse-grained source domain and fine-grained target domain*

Data with binary polarities can be more easily labeled than data with multiple emotion categories. However, it is a challenging task to do emotion tagging of comments when the source domain has binary sentiment categories and the target domain has fine-grained emotion categories.Figure 6 and Figure 7 show the comparison results on two groups of datasets with different sizes of target domain training data. More specifically, Table 5 provides the Accu@1,2,3 results under the training ratio 1/16 on QQ Entertainment dataset as target domain dataset.

From Figure 6 and Figure 7, we observe that the performance of the proposed probabilistic model P_CDCCET is not ideal. Even though P_CDCCET transfers knowledge from the source domain in the first level, it is too difficult to infer fine-grained categories from coarse-grained probability features in the second level.

When the ratio of the training data in the target domain is relatively low (e.g., below 1/8), E_CDCCET outperforms ETLR substantially. This indicates that our explicit model can take advantages of abundant data labeled with sentiment polarity to transfer knowledge and can help tag fine-grained emotion categories more accurately in the target domain which has a relatively small size. Unlike P_CDCCET which models the emotion category correlations probabilistically, E_CDCCET benefits from utilizing the explicit prior knowledge for modeling category correlation.

When the training ratio grows to 1/8 or higher, E_CDCCET loses advantages. This may be explained as that E_CDCCET utilizes the binary labeled data from the source domain in the first step of classification to remedy the lack of data in the target domain. Under higher training ratios, there turns to be fairly enough training data from the target domain for binary classification and the advantages gained from the auxiliary source domain data are no longer significant.

Statistical significance tests conducted for the Accu@1 results show that E_CDCCET outperforms ETLR with 0.95 confidence level when training ratio is lower or equal to 1/32 on all domain settings.

Table 5: The Accu@1,2,3 results with 1/16 training instances on QQ Entertainment dataset as the target domain with fine-grained sentiment categories and Sina Society dataset as the source domain but with binary sentiment polarity labels.

Methods	Accu@1	Accu@2	Accu@3
E_CDCCET	0.4527	0.5565	0.6812
P_CDCCET	0.3726	0.4844	0.5243
ETLR	0.4449	0.5301	0.6488

4.3.3 *Emotion Correlation*

We explicitly fix the matching relations between categories in E_CDCCET based on common sense. The corresponding matchings are as shown in Table 6.

For the probabilistic model P_CDCCET, the category correlations are trained instead of preassigned. Table 7 shows the learned correlations by normalizing the learned weights ν in Equation (8). They are sampled from one experiment when the source domain of Sina Society dataset has labels of 8 emotion categories while the target domain of QQ Entertainment has sentiment polarity as labels. Larger values indicate that the corresponding emotions are more likely to be positive, while smaller values indicate the opposite. "Happy" is the most positive emotion and "angry" is the most negative emotion. Besides, "touched", "sympathetic" and "surprised" are positive whereas "amused", "sad" and "anxious" are negative. These observations are quite consistent with common sense except for one mismatching. It explains the effectiveness of P_CDCCET model under the setting of fine-grained source domain and coarse-grained target domain.

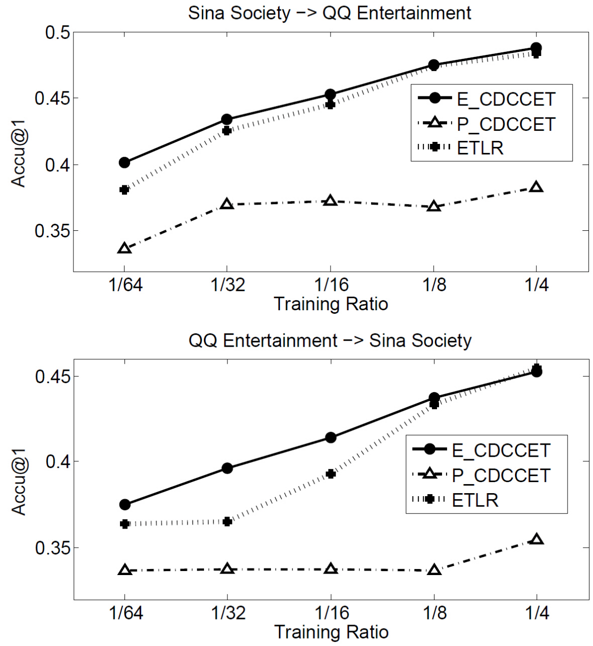

Figure 6: The Accu@1 results with coarse-grained source domain and fine-grained target domain on Chinese datasets.

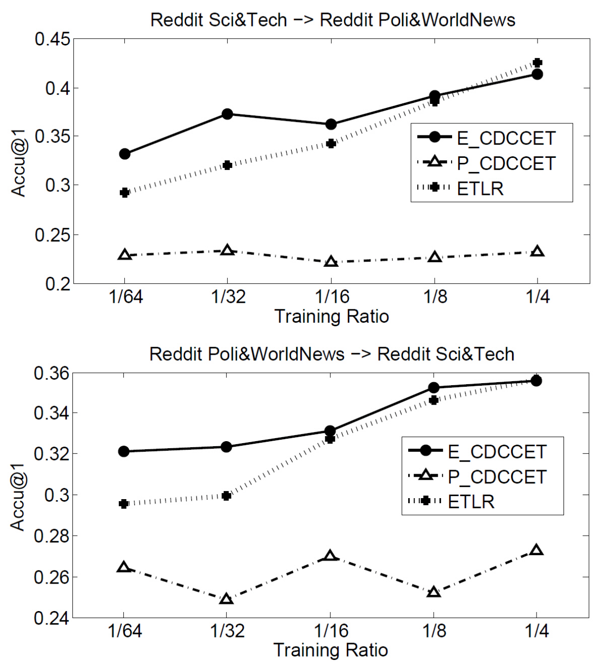

Figure 7: The Accu@1 results with coarse-grained source domain and fine-grained target domain on Reddit datasets.

Table 6: The pre-fixed emotion/polarity matchings.

Chinese Dataset		Reddit Dataset	
Positive	Negative	Positive	Negtive
Touched	Angry	Sympathetic	Angry
Sympathetic	Sad	Happy	Sad
Happy	Fervent	Surprised	Disgust
Amused	Bored		
Surprised	Anxious		
	Disgust		

Table 7: The normalized learned weights for emotion correlation on QQ Entertainment dataset.

Positive		Negative	
Emotion	Weight	Emotion	Weight
Touched	0.79	Amused	-0.19
Sympathetic	0.32	Angry	-0.73
Happy	1	Sad	-0.34
Surprised	0.56	Anxious	-0.36

5. CONCLUSIONS AND FUTURE WORK

This paper proposes a novel framework to address the task of predicting emotions for comments of cross-domain online news by modeling the relationship between different but related sets of emotion categories for the source and target domains. In particular, one method has been proposed for the task when the source and target domains share the same set of emotion categories and two methods have been proposed for the scenario when source and target domains use different categories. Our experimental results in both scenarios on four datasets demonstrate the effectiveness of the proposed approaches.

For possible future research, we plan to design a better text representation scheme by combining full text representation with feature selection techniques to avoid using only emotion terms. Furthermore, a more sophisticated modeling strategy for knowledge transfer may also improve the performance of cross-domain emotion tagging.

6. ACKNOWLEDGEMENTS

This work is partially supported by NSF research grants IIS-0746830, CNS-1012208, IIS-1017837 and CNS-1314688, and also partially supported by NSF of China research grants 61170184. This work is also partially supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

7. REFERENCES

[1] R. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *JMLR*, 2005.

[2] S. Bao, S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, and Y. Yu. Mining social emotions from affective text. *TKDE*, 2012.

[3] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, 2007.

[4] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *EMNLP*. ACL, 2006.

[5] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *ICML*, 2001.

[6] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[7] D. Das and S. Bandyopadhyay. Word to sentence level emotion tagging for bengali blogs. In *ACL*, 2009.

[8] D. Das and S. Bandyopadhyay. Sentence level emotion tagging on blog and news corpora. *JIS*, 2010.

[9] Z. Dong and Q. Dong. Hownet-a hybrid language and knowledge resource. In *NLP-KE*, 2003.

[10] P. Ekman, W. V. Friesen, and P. Ellsworth. Emotion in the human face: Guidelines for research and an integration of findings. 1972.

[11] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. *LREC*, 2006.

[12] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Springer Series in Statistics, 2001.

[13] J. Gao, H. Qi, X. Xia, and J. Nie. Linear discriminant model for information retrieval. In *SIGIR*, 2005.

[14] A. Genkin, D. Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 2007.

[15] B. Jakic and W. Weerkamp. Predicting sentiment of comments to news on reddit. 2012.

[16] A. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *NIPS*, 2002.

[17] L. Ku, Y. Liang, and H. Chen. Tagging heterogeneous evaluation corpora for opinionated tasks. *LREC*, 2006.

[18] T. Li, V. Sindhwani, C. Ding, and Y. Zhang. Knowledge transformation for cross-domain sentiment classification. In *SIGIR*. ACM, 2009.

[19] K. Lin, C. Yang, and H. Chen. What emotions do news articles trigger in their readers? In *SIGIR*, 2007.

[20] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *WWW*, 2007.

[21] T. Mullen and N. Collier. Sentiment analysis using support vector machines with diverse information sources. In *EMNLP*, 2004.

[22] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine learning*, 2000.

[23] S. Pan and Q. Yang. A survey on transfer learning. *TKDE*, 2010.

[24] B. Pang and L. Lee. Opinion mining and sentiment analysis. *FTIR*, 2008.

[25] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *ACL*, 2002.

[26] V. Sindhwani and P. Melville. Document-word co-regularization for semi-supervised sentiment analysis. In *ICDM*, 2008.

[27] R. Tokuhisa, K. Inui, and Y. Matsumoto. Emotion classification using massive examples extracted from the web. In *COLING*. ACL, 2008.

[28] P. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *ACL*, 2002.

[29] Q. Wang, L. Ruan, and L. Si. Adaptive knowledge transfer for multiple instance learning in image classification. In *AAAI*, 2014.

[30] C. Yang, K. Lin, and H. Chen. Emotion classification using web blog corpora. In *Web Intelligence*. IEEE, 2007.

[31] Y. Yang and X. Liu. A re-examination of text categorization methods. In *SIGIR*, 1999.

[32] D. Zhang, L. Si, and V. Rego. Sentiment detection with auxiliary data. *JIR*, 2012.

[33] H. Zhang, H. Yu, D. Xiong, and Q. Liu. HMM-based Chinese lexical analyzer ICTCLAS. In *SIGHAN workshop on Chinese language processing*, 2003.

[34] Y. Zhang, Y. Fang, X. Quan, L. Dai, L. Si, and X. Yuan. Emotion tagging for comments of online news by meta classification with heterogeneous information sources. In *SIGIR*, 2012.