Cluster-level Emotion Pattern Matching for Cross-Domain Social Emotion Classification

Endong Zhu
Sun Yat-sen University
Guangzhou, China
zhuend@mail2.sysu.edu.cn

Yanghui Rao*
Sun Yat-sen University
Guangzhou, China
raoyangh@mail.sysu.edu.cn

Haoran Xie
The Education University
of Hong Kong,
Hong Kong
hrxie2@gmail.com

Yuwei Liu
Sun Yat-sen University
Guangzhou, China
liuyw23@mail2.sysu.edu.cn

Jian Yin
Sun Yat-sen University
Guangzhou, China
issjyin@mail.sysu.edu.cn

Fu Lee Wang
Caritas Institute
of Higher Education,
Hong Kong
pwang@cihe.edu.hk

ABSTRACT

This paper addresses the task of cross-domain social emotion classification of online documents. The cross-domain task is formulated as using abundant labeled documents from a source domain and a small amount of labeled documents from a target domain, to predict the emotion of unlabeled documents in the target domain. Although several cross-domain emotion classification algorithms have been proposed, they require that feature distributions of different domains share a sufficient overlapping, which is hard to meet in practical applications. This paper proposes a novel framework, which uses the emotion distribution of training documents at the cluster level, to alleviate the aforementioned issue. Experimental results on two datasets show the effectiveness of our proposed model on cross-domain social emotion classification.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; Supervised learning by classification;

KEYWORDS

Emotion Detection; Cross-domain Classification; Clustering

1 INTRODUCTION

With the development of web 2.0, people tend to share opinions or emotions online. Among various types of online media, the online news website is an indispensable part which attracts many people to browse every day. Most of online news websites now provide a service which allows readers of an article to vote over several predefined emotion categories (e.g., happy, angry, and sad). The aggregation of these readers' emotional votes is termed as social emotion [2], and the research of social emotion classification on online news began with the SemEval-2007 tasks [9]. However, expressions that evoke a certain emotion may vary among different domains [8]. As a result, traditional emotion classification models require massive labeled data from various domains to guarantee the generalization ability of models. To overcome this shortcoming, cross-domain classification aims at utilizing the data in one specific domain to help build a robust model performing in other domains [3].

Most of the previous work of cross-domain classification focuses on eliminating the word-level difference [3][10]. One potential limit of the previous work is that these models require the distribution of features in different domains have a massive overlapping (i.e., some features should occur with a high frequency in all domains). This is because these models need to utilize overlapping features as intermediate information for building connections among domains. However, for online news, it may not always be convenient to meet this requirement because of a low overlapping of terminologies utilized in different news articles.

In light of these considerations, we propose a model based on cluster-level emotion distribution pattern matching for cross-domain social emotion classification. Specifically, we use the emotion distribution of labeled documents from the source domain and a small proportion of the target domain, and exploit the word distribution of all documents from the target domain to match and generate document clusters. The contributions of this paper mainly include: First, we propose a novel framework for cross-domain social emotion classification by assuming that similar documents in different domains tend to have similar emotion distributions. Second, we demonstrate that training a social emotion classifier at the weighted cluster level could enrich the contextual information and alleviate the negative effect caused by word-level differences in varied domains. Extensive experiments using two real-world datasets validate the effectiveness of our method.

*The Corresponding Author.

CIKM'17 , November 6–10, 2017, Singapore, Singapore

ACM ISBN 978-1-4503-4918-5/17/11. . . $15.00
https://doi.org/10.1145/3132847.3133063

2 RELATED WORK

Sentiment analysis, which deals with the computational tasks of annotating opinions and sentiments in text [7], is similar to emotion classification in natural language processing (NLP) areas. However, sentiment analysis mostly processes the data with coarse-grained (binary) categories (e.g., positive and negative), while emotion classification focuses on classifying the emotion data into fine-grained (multiple) categories (e.g., happy, sad and angry).

For cross-domain sentiment classification, Blizer et al. [4] introduced a structural correspondence learning (SCL) method by automatically inducing correspondences among features with some pre-selected pivot features. However, the SCL is developed for capturing the sentiment of writers rather than the evoked emotion of readers. For cross-context social emotion classification, Rao [8] proposed a contextual sentiment topic model (CSTM) from readers' perspective by distinguishing the context-independent themes from background themes and contextual themes. Although the CSTM takes the different feature distributions into consideration, it requires abundant features to gather enough statistics for topic modeling. Acharya et al. [1] showed that similar instances in the target domain are more likely to share the same class label and help detect possible differences across domains. However, their model is based on an optimization process and does not take the label distribution into consideration.

For cross-domain social emotion classification, Zhang et al. [10] made a compromise that a small proportion of documents in the target domain can also be labeled to provide more useful cross-domain information, and proposed a model based on reweighting source domain documents using a kernel function. Since this model is also concerned with the cross-domain social emotion classification task, we will use it as the main baseline in this study.

3 CLUSTER-LEVEL SOCIAL EMOTION CLASSIFICATION

3.1 Problem Definition

Given a source domain and a target domain, we use D_{src} to represent the set of documents in the source domain, D_{ltar} to represent the set of labeled documents in the target domain, and D_{utar} to represent the set of unlabeled documents in the target domain. For the i^{th} document d_i, let x_i represent the vector of word-level features (e.g., one-hot, term frequency, and TF-IDF) and y_i represent the vector of label distribution over every category. For example, a document may have a label distribution over four categories: {0.1, 0.2, 0.4, 0.3}, and 0.2 indicates that the document gets 20% reader votes over the second category. In addition, our model is trained via document clusters, so for the i^{th} cluster C_i, we use Y_i to represent its label vector. For simplicity, Y_i is estimated by the mean label distribution of all labeled documents in C_i.

3.2 Inner Domain Document Clustering

In this part, we conduct inner domain document clustering on both source and target domains. Within the same domain, the similarity between two documents is defined as their cosine similarity based on word-level vectors. For the source domain, we simply cluster all documents into k_1 clusters using a distance-based clustering algorithm (e.g., K-Means).

Due to the fact that documents in the target domain are divided into two sets D_{ltar} and D_{utar}, we first group documents in D_{ltar} into k_2 clusters using the distance-based clustering algorithm. The document set D_{ltar} is clustered separately rather than in conjunction with D_{utar} or D_{src}, because documents in D_{utar} are unlabeled while each cluster should contain at least one labeled document for model training, and documents' word-level features are different between D_{ltar} and D_{src}. Then for every document in D_{utar}, we compute the mean similarity between the document and every generated cluster's documents. In this way, we can add each document to the closest cluster in D_{ltar}. To alleviate the noise issue in clustering, we set a threshold θ to group the unlabeled target domain documents into the labeled target domain clusters when the similarity value is larger than θ.

3.3 Cross-Domain Cluster Matching

Since documents' word-level features are varied in different domains, we match and combine the aforementioned clusters across domains according to the similarity of emotion distributions. As a probability distribution for reader votes over all emotion labels, we use a symmetric form of Kullback-Leibler (KL) divergence [5] to measure the similarity between clusters, as follows:

$$KL(Y_i||Y_j) = \frac{1}{2}\sum_{k}^{dim}(Y_{ik}log\frac{Y_{ik}+\gamma}{Y_{jk}+\gamma} + Y_{jk}log\frac{Y_{jk}+\gamma}{Y_{ik}+\gamma}), \quad (1)$$

where Y_{ik} is the value of the k^{th} category in the label vector of C_i, dim is the number of categories, and γ is a smoothing term to avoid dividing by zero, which is set to 0.001 in our experiments. Here we use KL divergence because KL divergence is widely used to measure the similarity between two distributions (e.g., the probability distribution of emotion labels). In addition, our preliminary experiments demonstrated that KL divergence was robust and outperformed other measures like Euclidean distance and cosine similarity.

For a given pair of clusters, C_i and C_j, we compute the KL divergence between clusters using Y_i and Y_j. For every cluster in the target domain, we match it to the most similar cluster in the source domain and combine them into a new cluster, so the final number of clusters will be k_1. After combining clusters from source and target domains, we label unlabeled documents in each cluster using the mean vector of all other labeled documents' label vectors.

3.4 Cluster-based Model Training

Based on the combined cross-domain document clusters, we train a cluster-level classifier for social emotion prediction. To make an appropriate comparison with the cross-domain social emotion baseline model [10], we use the Logistic Regression with regulation term as the basic classifier. For every C_i, the cost function is given below:

$$\xi_i(\omega) = -\frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{k=1}^{dim} l_{jk}\log\frac{\exp(\omega_k \cdot x_j)}{\sum_{l=1}^{dim}\exp(\omega_l \cdot x_j)}, \quad (2)$$

where l is a binary indicator vector, s.t., $l_{jk} = 1$ if d_j is tagged with the k^{th} category, otherwise 0. N_i is the number of documents in C_i, and ω_k is the weight parameter of the k^{th} category. The total cost function is defined as a sum of every cluster cost function and one regularization term, and the formula is given as: $Cost(\omega) = \sum_{i=1}^{k_1} \lambda_i \xi_i(\omega) + R(\omega)$, where $R(\omega)$ is the L2 regularization term to avoid overfitting. Inspired by Li et al. [6], we assume that clusters with concentrated emotion distributions are more valuable for classification, because they are more likely to arouse an intensive emotion of readers. Based on this assumption, we assign different weights to clusters according to the standard deviation of clusters' label vectors. For C_i, its weight λ_i is given as: $\lambda_i = \sqrt{\frac{1}{dim}\sum_{k=1}^{dim}(Y_{ik} - \mu_i)^2}$, where μ_i is the mean value of C_i's labels. Then, the cost function is minimized by Gradient Descend method.

After the cluster-based model training process, we get the optimized weight parameter ω. Then, for every testing unlabeled document d_t in D_{utar}, we can predict the probability of every label as follows: $y_{tk} = \frac{\exp(\omega_k \cdot x_t)}{\sum_{l=1}^{dim} \exp(\omega_l \cdot x_t)}$, where y_{tk} is the value of the k^{th} category in the label vector of d_t, and x_t is the word vector of d_t.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

SemEval is an English dataset used in the 14th task of the 4th International Workshop on Semantic Evaluations [9], which is a short-text dataset that consists of news headlines and reader ratings over 6 emotions. The emotion labels include anger, disgust, fear, joy, sad and surprise. The 246 training documents are considered as the source domain, and the 1000 testing documents form the target domain. The source domain and the target domain documents are very different in word distributions due to different published time of news. Particularly, the proportion of number of words that occur in both domains to the size of vocabulary is only 19.5%, which indicates that the overlapping in word distributions is quite small for the training and testing documents.

ChinaNews[1] is a Chinese dataset that contains reader ratings over 8 emotions toward news articles on ChinaNews website (www.chinanews.com), which is a long-text dataset from April, 2010 to August, 2011. The emotion labels include moved, sympathy, boring, anger, funny, sad, delighted and not-interested. The first 1913 documents are from the law news domain, and the rest 1571 documents are from the economic news domain. We use the law news as the source domain and the economic news as the target domain.

We employed two evaluation metrics as indicators of performance: $Acc@1$ [10] and AP [8]. The coarse-grained metric $Acc@1$ stands for the accuracy at top 1. $Acc@1$ is computed by dividing the number of correctly predicted documents by the total number of documents. The fine-grained metric AP is the average Pearson's correlation coefficient over all documents. For each document, AP measures the correlation between the predicted probabilities and the truth votes over all emotion categories.

4.2 Results and Analysis

In this part, we estimate the effectiveness of our proposed model and baselines. The following models are compared:

- **Single Domain Logistic Regression (SDLR):** A naïve method which directly uses labeled documents in the target domain to train a Logistic Regression classifier.
- **Single Domain Cluster-level Logistic Regression (SDCLR):** A method which first clusters all documents in the target domain by word distributions, and then train a Logistic Regression classifier with weighted clusters.
- **Cross-Domain Cluster-level Emotion Pattern Matching (CDEPM):** The proposed model which utilizes the information from the source domain through cross-domain cluster matching, and the model also uses Logistic Regression as the basic classifier.
- **Cross-Domain Emotion Tagging with Joint Probabilities (CDET_J):** A method proposed by Zhang et al. [10] which reweights the source domain data based on joint probabilities. This model is implemented using Logistic Regression as the basic classifier.
- **Structral Correspondence Learning (SCL):** A method proposed by Blitzer et al. [4]. We implemented this model with Logistic Regression. The pivot features are chosen as the top 1% frequently co-occurred terms.
- **Contextual Sentiment Topic Model (CSTM):** A topical model proposed by Rao [8], which distinguishes the context-independent themes specifically.

For each dataset, we set the ratio of data labeled in the target domain for training to 1/64, 1/32, 1/16, 1/8, 1/4, and 1/2, and the size of validation sets are also proportional to this ratio. Validation sets will be used for the selection of parameters θ, k_1 and k_2 in the proposed model, as well as the parameter setting of baseline models. According to sufficient experimental results on a subset of our employed datasets, we range θ from 0.25 to 0.35. We will also discuss the influence of k_1 and k_2 later in this part. To well represent word features of both short-text and long-text corpora, the term frequency is used in generating document vectors. Gradient Descent method is employed to estimate parameters for all models that use Logistic Regression as the basic classifier. For each setting of training ratios, we separately run the models for 10 times and report the mean values of two evaluation metrics. The results are reported in Tables 1-2.

From the performance of SDLR, SDCLR and CDEPM, we can evaluate the effectiveness of each module of our model. We can observe that SDLR performs poorly on every setting due to the inadequate word-level information. By comparing SDLR and SDCLR, we can observe that training in the form of weighted clusters could improve the performance. This is because clusters may contain some unlabeled documents from the target domain, and these documents will enrich the set of utilized words. As for CDEPM, it outperforms other models in most settings, which demonstrates the effectiveness of

[1]The dataset is available in public at www.dropbox.com/sh/hkch4b5h6pavqi5/AACpz4twsOmBrFb1AuLTOnL_a?dl=0

Table 1: Results on SemEval with different training ratios (Rt) in terms of two metrics (Met)

Rt	Met	SDLR	SDCLR	CDEPM	CDET_J	SCL	CSTM
1/64	$Acc@1$	0.191	0.263	**0.371**	0.340	0.333	0.317
	AP	0.234	0.252	0.323	**0.326**	0.311	0.303
1/32	$Acc@1$	0.235	0.266	**0.380**	0.347	0.349	0.358
	AP	0.260	0.277	**0.335**	0.291	0.304	0.312
1/16	$Acc@1$	0.260	0.328	**0.390**	0.367	0.358	0.361
	AP	0.267	0.292	**0.351**	0.347	0.335	0.334
1/8	$Acc@1$	0.341	0.362	**0.402**	0.397	0.389	0.377
	AP	0.355	0.328	0.379	**0.393**	0.372	0.355
1/4	$Acc@1$	0.399	0.415	**0.423**	0.411	0.402	0.393
	AP	0.386	0.378	0.393	**0.408**	0.381	0.371
1/2	$Acc@1$	0.422	0.424	**0.448**	0.436	0.427	0.404
	AP	0.372	0.380	**0.404**	0.400	0.393	0.387

Table 2: Results on ChinaNews with different training ratios (Rt) in terms of two metrics (Met)

Rt	Met	SDLR	SDCLR	CDEPM	CDET_J	SCL	CSTM
1/64	$Acc@1$	0.535	0.544	**0.564**	0.548	0.551	0.561
	AP	0.541	0.553	0.601	0.594	0.585	**0.607**
1/32	$Acc@1$	0.547	0.551	**0.567**	0.552	0.555	0.563
	AP	0.587	0.592	0.607	0.603	0.591	**0.611**
1/16	$Acc@1$	0.550	0.557	**0.571**	0.563	0.560	0.568
	AP	0.571	0.580	**0.618**	0.617	0.607	0.614
1/8	$Acc@1$	0.553	0.564	**0.587**	0.579	0.571	0.580
	AP	0.592	0.613	**0.631**	0.628	0.622	0.624
1/4	$Acc@1$	0.562	0.570	**0.591**	0.588	0.582	0.587
	AP	0.609	0.632	**0.647**	0.646	0.624	**0.647**
1/2	$Acc@1$	0.614	0.642	**0.663**	0.651	0.648	0.656
	AP	0.659	0.692	**0.698**	0.685	0.688	0.693

matching cross-domain clusters by emotion label distribution. In addition, it is notable that CDEPM has predominant performance when only a small proportion (e.g., 1/64, 1/32, and 1/16) of the target domain data is used. As for the baseline models, SCL's performance improves as the increase of training ratios. This is might because pivot features of high occurrence frequency can represent the correspondence of domains with a sufficient proportion of labeled target domain documents. CSTM's performance on ChinaNews dataset is much better than that on SemEval. This is because CSTM requires abundant features to gather enough statistics. To statistically evaluate the differences between CDEPM and CDET_J, we conduct significance tests on them using t-distribution with a 0.95 confidence level. From the test result, we observe that our model CDEPM outperforms CDET_J with an obvious difference in most cases.

As a baseline model without the supervision of any labeled data in the target domain, a Logistic Regression classifier using term frequencies and labels in the source domain as input and output is also implemented. For this method, the results on SemEval are 0.337 and 0.311 in terms of $ACC@1$ and AP, respectively. As for ChinaNews, the $ACC@1$ and AP values of the above method are respectively 0.546 and 0.571. These unsatisfactory results validate that the word distributions of documents in the source and target domains are very different for both datasets.

To demonstrate the influence of parameters on our model CDEPM, we vary k_1 and k_2 to further examine their effects. We independently change the ratio of k_1 to the number of source domain documents (k_1 ratio) and the ratio of k_2 to the number of target domain documents (k_2 ratio) with other parameters fixed. We find that k_1 ratio has a more obvious influence than k_2 ratio on results of both datasets. This is might because that k_1 directly determines the final number of cross-domain clusters, and thus has a more direct influence on the training process. In addition, we find that the best value of k_1 ratio ranges from 0.9 to 1, while the best value of k_2 ratio ranges from 0.85 to 0.95. This indicates that the documents' word distributions in our employed datasets are extremly varied.

5 CONCLUSIONS

In this paper, we propose a model named CDEPM for cross-domain social emotion classification based on cluster-level emotion distribution pattern matching. The experimental results clearly demonstrate the effectiveness of our model in terms of $ACC@1$ and AP on both short-text and long-text datasets. In future work, we plan to reduce the complexity of our model by exploiting sparse coding of clustering algorithms to improve its efficiency on large-scale datasets. Extensive experiments over a dataset across more than two domains will also be conducted to evaluate the generalization ability of different models. Furthermore, we plan to employ other evaluation metrics (e.g., running time) to compare the proposed model with baselines comprehensively.

ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China (61502545, 61472453, U1401256, U1501252, U1611264), Research Grants Council of Hong Kong Special Administrative Region, China (UGC/FDS11/E03/16), and the Internal Research Grant (RG 66/2016-2017) of The Education University of Hong Kong.

REFERENCES

[1] A. Acharya, E. R. Hruschka, J. Ghosh, and S. Acharyya. 2014. An optimization framework for combining ensembles of classifiers and clusterers with applications to nontransductive semisupervised learning and transfer learning. *ACM Transactions on Knowledge Discovery from Data* 9, 1 (2014), 1–35.

[2] S. Bao, S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, and Y. Yu. 2012. Mining social emotions from affective text. *IEEE Transactions on Knowledge and Data Engineering* 24, 9 (2012), 1658–1670.

[3] J. Blitzer, M. Dredze, F. Pereira, et al. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*. 440–447.

[4] J. Blitzer, R. McDonald, and F. Pereira. 2006. Domain adaptation with structural correspondence learning. In *EMNLP*. 120–128.

[5] W. De Smet and M. F. Moens. 2009. An aspect based document representation for event clustering. In *CLIN*.

[6] X. Li, Y. Rao, Y. Chen, X. Liu, and H. Huang. 2016. Social emotion classification via reader perspective weighted model. In *AAAI*. 4230–4231.

[7] B. Pang, L. Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2, 1–2 (2008), 1–135.

[8] Y. Rao. 2016. Contextual sentiment topic model for adaptive social emotion classification. *IEEE Intelligent Systems* 31, 1 (2016), 41–47.

[9] C. Strapparava and R. Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *SemEval*. 70–74.

[10] Y. Zhang, N. Zhang, L. Si, Y. Lu, Q. Wang, and X. Yuan. 2014. Cross-domain and cross-category emotion tagging for comments of online news. In *SIGIR*. 627–636.