A Common Subspace Construction Method in Cross-Domain Sentiment Classification*

Yuhong Zhang, Xu Xu, Xuegang Hu

School of computer science and information technology
Hefei University of Technology, 230009, China
zhangyh@hfut.edu.cn, xu20085889@163.com, jsjxhuxg@hfut.edu.cn

***Abstract*—In this paper, we study the problem of domain adaptation in sentiment classification. Many existing approaches reduce the gap by extracting domain-independent topics. However these methods couldn't cope with features which have different sentiments in different domains. To solve this problem, a common subspace construction method (CSC) is proposed in our paper. Firstly, the consistency of features' sentiment orientation in different domains is introduced to identify the common subspace. Then, domain-dependent features will be projected to this subspace. Empirical studies on benchmark tasks of sentiment analysis validate our assumption and demonstrated significant improvement of our method over competing ones in classification accuracies.**

Keywords-cross-domain; sentiment classification; common subspace

I. INTRODUCTION

The explosion of unlabeled text on the web such as microblog, shopping reviews makes cross-domain sentiment classification an important and challenging task [1-7]. Cross-domain sentiment classification aims to adapt the classifier trained on the labeled source domain for the unlabeled target domain. Normally, the two domains are similar but not identical.

Many existing approaches [3], [5], [6] consider that the classifier trained on the source domain is no longer applicable to the target domain because there are amount of features which are not included in the trained classifier. For example, in the electronics domain, "*durable*" is used to express positive sentiment, and "*short-battery*" is used to express negative sentiment. While "*durable*" and "*short-battery*" are not used in the books domain. In the same way, features such as "*thriller*" are not appearing in the electronics domain. Therefore, the sentiment classifier trained on the electronics domain will not work well in the books domain. This problem is called as the mismatch of domain[7].

To solve this problem, existing approaches like SCL[3][4], SFA[5] tried to extract domain independent topics to construct subspace and train a classifier in this subspace. These topics are the transformation of the original features, so that, domain-dependent features in the target domain could be mapped to features in the source domain. However, in practice, these methods couldn't cope with the features which have different sentiments in different domains[8], because a topic only can express one kind of sentiment, and these features will be wrongly trained in source domain.

In this paper, we aim to construct a more accurate common subspace, and then train classifier for the target domain based on the common subspace. Given a labeled source domain *S* and an unlabeled target domain *T*, we can compute the *SO* (sentiment orientation) of features in source domain using methods CPD (Categorical Proportional Difference) [9] or OR (odds ratio) [10]. Then we predict the sentiment orientation of these features in the target domain base on the relationship of co-occurrence, and the common subspace is constructed according to the consistency of sentiment orientation between different domains. Thus, domain-dependent features can be filtered. Lastly, all of domain-dependent features are projected to common subspace to solve the domain mismatch.

The rest of the paper is organized as follows. In the next section, we first review some related works about cross-domain sentiment classification, especially the common subspace construction methods. The details of our method are presented in section 3. Section 4 describes experimental results. Finally, section 5 concludes this paper.

II. RELATED WORK

Recently cross-domain sentiment classification amuses wide attention. Most current approaches resolve the mismatching of domain through identifying a subspace where data in the source and target domain are similarly distributed. SCL[3] is a representative one. It tried to get the mapping matrix from all features to a common feature subspace. Domain-dependent features are then transferred through the mapping matrix. SFA[5] used some domain-independent features as a bridge to construct domain-special feature clusters. In SST [11] method, it used the related features from source domain to expand vectors in a binary classifier at training and testing times. HeMap[12] tried to use spectral transformation construct common subspace. HFA[13] tried to augment the original feature space, and then, project all of features to a subspace. Transfer-PLSA [6] extracted topics between different domains, so that domain-dependent features can be transferred across different domains. Another work [7] tried to identify a subspace where data in source and target domains are similarly and discriminatively distributed.

But all of these methods ignore that a feature which map have different sentiments in different domains. Different from above methods, we select the domain-independent features to construct common subspace whose sentiment orientation are consistent in the source and target domain. And then we project all of features to the subspace. In our method, the domain-dependent feature will get different projection values on the subspace according to the domain.

III. THE PROPOSED APPROACH

Before proposing our method in detail, we first give some definitions and problem setting in this paper.

Definition 1 (domain-independent feature): Domain-independent feature should not only occur frequently in both domain, but also has the some sentiment orientation in the source and target domains, such as "*excellent*" and "*worthless*".

Definition 2 (domain-dependent feature): Domain-dependent features occur in only one domain or occur with both domains but associated with different sentiment orientation. For example, feature "*boring*" appears in books frequently, but we can't find it in kitchen. Feature "*end-up*" occurs in both books and kitchen, but associated with different sentiment orientation in the two domains.

Problem setting: Assuming that there is a source domain denoted by *S* in which data are labeled and a target domain denoted by *T* in which data are unlabeled. Labels are denoted by $y \in \{1,-1\}$, where 1 represents positive and -1 represent negative. Our object is to get a common feature subspace, and a classifier trained on it with a high accuracy for target domain. For the sake of apprehension, we list the symbols used in this paper in Table 1.

TABLE I. SUMMARY OF SYMBOLS USED IN OUR ALGORITHM

Symbol	Description
D	*D* is a domain variable. It's value can be source domain *S* or target domain *T*
x_{D_i}	The *i*-th text in domain *D*
y_{D_i}	The label of text x_{D_i}
$TF(f, x_{D_i})$	The times of feature *f* appearing in text x_{D_i}
$SO_D(f)$	The sentiment orientation of feature *f* in domain *D*
$h_{D+}(f)$	The proportion of the feature *f* occurring in domain *D* positive texts, the same as $h_{D-}(f)$
F_C	The set of features in common subspace
nc	The number of domain-independent features in F_C
F_S	The set of features in source domain
F_T	The set of features in target domain

A. Common Subspace Construction

In this subsection we will select the domain-independent features to construct common subspace according their sentiment orientation in both domains. Normally, we can use the difference of frequency in positive text and negative text to represent the sentiment orientation of features, such as CPD (Categorical Proportional Difference) [9] and OR (odds ratio) [10]. In this paper we use CPD to calculate the sentiment orientation of features, its formula is shown in Eq. (1):

$$SO_D(f) = \frac{h_{D+}(f) - h_{D-}(f)}{h_{D+}(f) + h_{D-}(f)} \tag{1}$$

Where $h_{D+}(f)$ represent the proportion of the feature *f* occurring in positive texts in domain D, $h_{D+}(f)$ can be obtained using Eq. (2), and $h_{D-}(f)$ can be computed in the similar way.

$$h_{D+}(f) = \frac{\sum_{y_{D_i}>0} FP(f, x_{D_i}) * y_{D_i}}{\sum_{y_{D_i}>0} y_{D_i}} \tag{2}$$

Where $FP(f, x_{D_i})$ denotes whether feature *f* appears in the document x_{D_i} or not, its value is 1 when *f* appears in x_{D_i}, otherwise its value is 0.

Here we could get $SO_S(f)$, the sentiment orientation of features in the source domain according to Eq. (1), but the sentiment orientation in the target domain $SO_T(f)$ is unobtainable because of the lack of reviews label y_{T_i}. In order to compute $SO_T(f)$, we use $SO_S(f)$ to calculate $h_{T+}(f)$, as Eq. (3) shows. And the same as $h_{T-}(f)$:

$$h_{T+}(f) = \frac{\sum_{p_s(f_i)>0, f_i \neq f} PMI_T(f, f_i) * p_s(f_i)}{\sum_{p_s(f_i)>0, f_i \neq f} p_s(f_i)} \tag{3}$$
$$|p_s(f_i)| > \alpha \quad f, f_i \in F_S \cap F_T$$

We compute $PMI_T(f, f_i)$ as the point-wise mutual information between feature f and feature f_i. In which $f, f_i \in F_S \cap F_T$. In order to guarantee the discriminative of f_i, we require $SO_S(f)$ larger than α. It is worth to note that $PMI_T(f, f_i)$ is computed only based on the target domain not on the union of source and target domains, it denotes the similarity between the two features in target domain.

Now, according to Eq. (1) and (3), $SO_T(f)$ could be predicted. In view of errors in the predicting process, we sort $SO_T(f_j)$ and select the top-*k* positive and negative features to ensure their reliability and discrimination. Then we will select the features, whose $SO_S(f)$ is consistent with $SO_T(f)$ as common features to construct subspace. The formally representation is shown in Eq. (4).

$$F_c = \{f_c^i \mid SO_s(f_c^i) * SO_T(f_c^i) > 0\} \tag{4}$$

Let's illustrate the process with an example. Table 2 shows the selection of domain-independent features. Feature "*out*" will be filtered because we couldn't ensure it's a sentiment word in the target domain according to its weak predicted value; whereas feature "*end-up*" is also discarded because whose inconsistency between $SO_S(f)$ and $SO_T(f)$, although their value is high. Contrarily, "*excellent*" and "*worthless*" are selected as domain-independent features to construct the common subspace.

TABLE II. AN EXAMPLE OF DOMAIN-INDEPENDENT FEATURE SELECTION IN THE TASK OF BOOKS- KITCHEN

Features	$SO_S(f)$	$SO_T(f)$
excellent	0.84	0.53
worthless	-0.36	-0.83
end-up	0.11	-0.35
out	-0.02	-0.09

B. Frequency Expansion Method

The common subspace has been selected in the previous subsection. However, the classifier trained on the subspace may not achieve a good result, because domain-dependent features are lost in the selection process. To overcome this problem, we propose a frequency expansion method to solve the mismatching of domains.

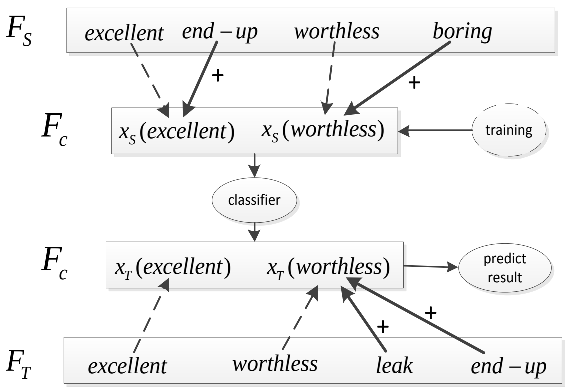

Fig. 1. illustration of our algorithm

The Figure 1 illustrates the main idea of our algorithm. The F_S and F_T in Figure 1 represent the feature set of source domain (books) and target domain (kitchen) respectively. The source domain contain words "*excellent*", "*end-up*", "*worthless*" and "*boring*", and the target domain contain words "*excellent*", "*worthless*", "*leak*" and "*end-up*". According to the previous section, the common subspace contains "*excellent*" and "*worthless*". However, if a review only contains "*end-up*", "*boring*" or "*leak*", the vector of the review in the common subspace will be <0,0>. To overcome this problem, we project these features to their relevant domain-independent features. Namely that "*leak*" is projected to worthless, and "*end-up*" is projected to "*excellent*" and "*worthless*" in source domain and target domain respectively. Moreover, we project all of domain-dependent features appearing in the review together, instead of considering each feature individually. The process of projection could be described as Eq. (5) and (6).

$$PMI\left(f_c, x_{D_i}\right) = \frac{\sum_{f_i \notin F_c \ f_i \in F_D} TF(f_i, x_{D_i}) PMI_D(f_c, f_i)}{\sum_{f_i \notin F_c \ f_i \in F_D} TF(f_i, x_{D_i})} \quad (5)$$

In the Eq. (5), $PMI\left(f_c, x_{D_i}\right)$ denotes the value of projection from the domain-dependent features occurring in x_{D_i} to f_c. According to the definition, given a review x_{D_i}, the common feature f_c will have a high score if there are many features co-occurring with f_c frequently in x_{D_i}. Moreover, we use $PMI_S(f_c, f_i)$ and $PMI_T(f_c, f_i)$ separately, so that the feature which has different sentiment orientation in different domains could be projected to different word. Then, the project function is normalized in Eq. (6).

$$s\left(f_c, x_{D_i}\right) = \begin{cases} \dfrac{PMI\left(f_c, x_{D_i}\right)}{\max\limits_{f_c}(s(f_c))} & PMI\left(f_c, x_{D_i}\right) > 0 \\ 0 & PMI\left(f_c, x_{D_i}\right) \le 0 \end{cases} \quad (6)$$

Next, we model a review x_{D_i} on the common subspace, and then represent the review by a vector $<v_1 ... v_{nc}>$, where the v_j is the sum of $TF(f_c^j, x_{D_i})$ and the project score $s\left(f_c^j, x_{D_i}\right)$. The vector of x_{D_i} can be denoted as follows:

$$x_{D_i} = < TF(f_c^1, x_{D_i}) + s\left(f_c^1, x_{D_i}\right), \ldots, TF(f_c^{nc}, x_{D_i}) + s\left(f_c^{nc}, x_{D_i}\right) > \quad (7)$$

Where *nc* denote the number of domain-independent features in F_C. At last, we train a binary classifier on the subspace to predict the sentiment orientation of target domain reviews.

Algorithm Common subspace construction method for cross domain sentiment classification (CSC)

Input: labeled source domain $S = \{(x_{S_i}, y_{S_i})\}_{i=1}^{n_s}$, unlabeled target domain $T = \{x_{T_i}\}_{i=1}^{n_T}$, the value of k and α

Output: classifier C for T

1. Calculate $SO_S(f)$ in the domain S and predict features' sentiment orientation in the domain T
2. Apply the consistency criteria to select domain-independent features, and construct the common subspace F_C.
3. Calculate the project functions value $s\left(f_c, x_{D_i}\right)$ from review x_{D_i} to domain-independent feature f_c
4. Represent the review x_{D_i} on the common subspace by the following vector $\{x_{S_i}\}$ and $\{x_{T_i}\}$.
5. Return a classifier C, trained on $\{x_{S_i}\}$, and test C on $\{x_{T_i}\}$

IV. EXPERIMENTS AND EVALUATIONS

A. Data Sets and Baseline Algorithms

We evaluate our algorithm on the RevData[4], which has be widely used in many cross domain sentiment classification methods . The dataset consists of product reviews collected from four different domains of amazon.com- DVD(D), kitchen(K), books(B), electronics(E). Each domain includes 1000 positive and 1000 negative reviews. On this data set we could construct 12 cross-domain tasks denoted by D->B,E->B,K->B, e.g., the word before the arrow means the source domain and the word after the arrow means the target domain.

Next, we describe some baselines with which we compare in our method.

No-Tra: We train the classifier on the source domain apply it to the target domain directly.

SCL-MI: This method augmented original features with linearly transformed topics.

SFA: This method selected the features using mutual information between feature and domain label as common features, and then introduced the spectral clustering to align the words from different domain to help bridge the gap between them.

None Frequency Expansion (CSC-NFE): It is a variant of CSC. It uses the frequency of features to construct the vector of reviews in common subspace, not considering the project score.

CSC-MI: It is also a variant of CSC, it uses MI in source domain to select domain-independent features (like SCL-MI).

CSC: our algorithm.

B. Parameter Setting

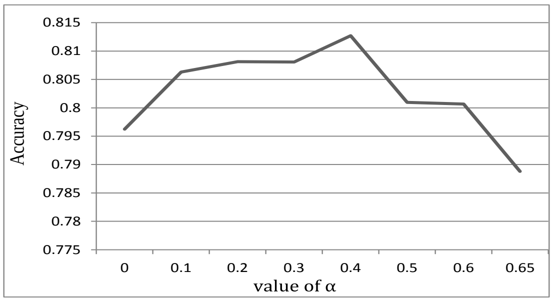

Fig. 2. Accuracy varying with α

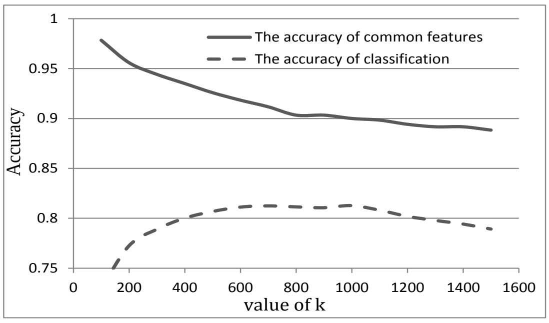

Fig. 3. Accuracy varying with *k*

There are two main parameters in our algorithm, the value of α in selecting common features set F_C and the number of *k*. So we conduct an experiment to select the best values as shown in Figure 2 and 3. The accuracy in these Figures represents the average results of 12 cross-domain tasks. Figure 2 describes the relationship of α and the accuracy of cross domain classification, and we can see from it that our algorithm performs best when α falls in 0.4, because when the α is lower than 0.4, the accuracy of these features is low too, and when the α is larger than 0.4, we can’t select enough features, both of them will affect the process of features’ sentiment orientation prediction in target domain.

In Figure 3, the solid line represents the accuracy of domain-independent features in subspace and the dashed line represents the result of classification. From Figure 3, we can see that the more features are selected, the lower accuracy of subspace we will get, at the same time, the classification accuracy experiences a process from low to high and then high to low. Ultimately, we set *k*=1000 in our experiment.

C. The effectiveness of common Subspace

In this section, we compare the accuracy of domain-independent features in subspace to demonstrate the effective of our feature selection method. The accuracy of subspace is crucial in our algorithm, one reason is that the common subspace can be regard as bridge to link the different domain-dependent features, and the other reason is the classifier will be trained on the subspace. Existing approaches only select initial domain-independent features to assist extracting domain-independent topics. For example, SCL-MI selected features with high sentiment orientation in the source domain, SFA selected features with lower mutual information between features and domain labels.

The Table 3 shows the comparison results of different methods. For the sake of simplicity, all of numbers are the average value of 12 tasks. In Table 3, NF(Number of Features) denote the number of domain-independent features selected by different methods, NS(Number of Source domain-dependent features) denote the features which have been selected as domain-independent features but not exist in the target domain, AS (Average of absolute Sentiment orientation in the target domain) denotes the discrimination of features, AC(accuracy of features) denotes the accuracy of domain-independent features after removed NS.

From the Table 3, we can observe that there are 4623 features in the source domain on average, and nearly half of them are domain-dependent features. In order to filter out these domain-dependent features, SCL-MI selected 600 features according to MI in the source domain, so it has the highest value of AS, but it also selected some domain-dependent features, as evidenced by its high value of NS, low value of AC. SFA selected features appearing in both domain frequently, so it has a higher AC than SCL-MI and the value of NSSF is 0, but its AS value is only 0.25, lower than all of method, because it tends to select neutral features as domain-independent features, meanwhile, SFA also can't filter out the features which associated different sentiments in different domains. Contrarily, our method has a high value of AS and the best AC, because we could guarantee the discrimination of features by selecting the top-*k* predicted sentiment orientation, and then, we also filter out the domain-dependent features according the consistency of features' orientation in different domain. Lastly, we selected 652 features as domain-independent features on average.

TABLE III. ACCURACY COMPARISON OF DIFFERENT METHODS IN DOMAIN-INDEPENDENT FEATURE SELECTION

method	NF	NS	AS	AC
F_S	4623	643	0.36	0.52
SCL-MI	600	98	0.51	0.69
SFA	600	0	0.25	0.75
CSC	652	0	0.38	0.87

D. Accuracy in Cross-Domain Classification

TABLE IV. ACCURACY COMPARISON OF DIFFERENT METHODS IN CROSS-DOMAIN CLASSIFICATION (%)

	No-Tra	SCL-MI	SFA	CSC-MI	CSC-NFE	CSC
D-K	77.2	81.4	80.75	80.3	81.15	**84.75**
D-B	77.5	**79.7**	77.5	**79.7**	78.9	79.05
D-E	75.45	74.1	76.7	79.9	79.6	**83.3**
K-D	72.4	76.9	76.95	73.6	78.15	**80.25**
K-B	71.8	68.6	74.8	76.7	76.15	**77.55**
K-E	79	**86.8**	85.05	84.95	84.2	84.6
B-D	70	75.8	**81.35**	76.5	79.15	80.75
B-K	67	78.9	78.8	80.7	79.6	**83**
B-E	61.15	75.9	72.5	75.1	75.65	**79.7**
E-D	69.7	76.2	77.15	74.6	77.15	**80.35**
E-K	82.2	85.9	**86.75**	85.55	84.95	85.65
E-B	68	75.4	75.65	71.45	74.4	**76.3**
Ave	72.2	77.96	78.66	78.2	79	**81.27**

In this section, we compare results of our algorithm against baselines to demonstrate the effect of our method on the performance of cross-domain classification. In these methods, SCL-MI, SFA and CSC-MI couldn't filter out the features which are associated with different sentiment orientations in different domains, and the CSC-NFE haven't utilized domain-dependent features. In the Table 4, the digitals which significantly greater than others are bold tagged. Firstly, by contrasting CSC-NFE with SCL-MI, SFA and CSC-MI, we can observe that the accuracy of classification could be improved greatly by eliminating features which are associated with different sentiment orientations in different domains. In the other words, these features will largely drop the performance of cross-domain classification. Then, by contrasting CSC-NFE and CSC, we can observe that the accuracy of classification could be improved by project domain-dependent features to common subspace, it demonstrates the effective of frequency expand method. At last, our final results perform significantly better than all the other algorithms', from this point, we can see that our algorithm could project domain-dependent features to subspace perfectly. These results demonstrate that our algorithm is more effective.

V. CONCLUSIONS

In this paper, we propose the feature which belong to domain-dependent features and associates with different sentiments in different domains. Based on selecting a higher accuracy of common subspace according to the consistency of sentiment orientation in different domains, we settle the problem by projecting domain-dependent features to the common subspace. Our experimental results show the effective of our approach.

In the future, we plan to extend our work in the following directions: 1) Constructing subspace on unbalanced data set. 2) Studying the effect of neutral features on cross-domain classification.

ACKNOWLEDGMENT

This work is supported in part by the National Natural Science Foundation of China (NSFC) under grants 61305063 and 61273292, Doctoral Fund of Ministry of Education of China under grant 2013JYBS0632.

REFERENCE

[1] S. J. Pan, and Q. Yang, "A survey on transfer learning," Knowledge and Data Engineering, 22(10), pp.1345–1359, August 2010.

[2] H. Daume III, and D. Marcu, "Domain adaptation for statistical classifiers," Journal of Artificial Intelligence Research, 26, pp.101-126, 2006.

[3] J. Blitzer, R. McDonald, and F. Pereira, "Domain Adaptation with Structural Correspondence Learning," Proceedings of Empirical Methods in Natural Language Processing. Sydney, Australia. pp. 120-128, 2006

[4] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," Proceedings of the 45th Association for Computational Linguistics. Prague, Czech Republic. pp. 440-447, 2007.

[5] S. J. Pan, X. C. Ni, J. T. Sun, Q. Yang, and Z. Chen. "Cross-Domain Sentiment Classification via Spectral Feature Alignment," Proceedings of the 19th International World Wide Web Conference. Raleigh, North Carolina, USA: ACM. pp. 751-760. 2010.

[6] K. Liu, and J. Zhao, "Cross-Domain Sentiment Classification Using a Two-Stage Method," Proceedings of the 18th ACM conference on Information and knowledge management. Hong Kong, China. pp.1717-1720, 2009.

[7] Y. Shi, F Sha, "Information-theoretical learning of discriminative clusters for unsupervised domain adaptation," Proceedings of the 29th International Conference on Machine Learning. Edinburgh, Scotland, UK. 2012.

[8] P. Biyani, C. Caragea, P. Mitra, C. Zhou, J. Yen, Greta E. Greer, and K. Portier, "Co-training over domain-independent and domain-dependent features for sentiment analysis of an online cancer support community," International Conference on Advances in Social Networks Analysis and Mining. Ontario, pp.413-417, 2013.

[9] M. Simeon, and R. Hilderman, "Categorical proportional difference: A feature selection method for text categorization," The Australasian Data Mining Conference. pp. 201–208, 2008.

[10] M. Whitehead, and L. Yaeger. "Building a General Purpose Cross-Domain Sentiment Mining Model," Proceedings of the Computer Science and Information Engineering. Los Angeles, CA. pp. 472-47, 2009.

[11] D. Bollegala, D. Weir, and J. Carroll. "Cross-Domain Sentiment Classification using a Sentiment Sensitive Thesaurus," IEEE Transactions on Knowledge and Data Engineering, 25(8), pp. 1719-1731. 2013.

[12] X. Shi, Q. Liu, and W. Fan, "Transfer learning on heterogenous feature spaces via spectral transformation," In ICDM, Sydney, NSW, pp. 1049-1054, 2010.

[13] L. Duan, D. Xu, and Ivor W.Tsang. "Learning with Augmented Features for Heterogeneous Domain Adaptation," Proceddings of the 29th Internation Conforence on Machine Learning, Edinburgh, Scotland, UK, 2012.