Improving generalization ability of Instance-transfer Based Imbalanced Sentiment Classification of Turn-Level Interactive Chinese Texts

Tian F, Wu F, Fei X, Shah N, Zheng Q, Wang Y

Tian, F, Wu, F, Fei, X, Shah, N, Zheng, Q & Wang, Y 2019, 'Improving generalization ability of Instance-transfer Based Imbalanced Sentiment Classification of Turn-Level Interactive Chinese Texts', Service Oriented Computing and Applications, vol. 13, no. 2, pp. 155-167.
https://doi.org/10.1007/s11761-019-00264-y

DOI 10.1007/s11761-019-00264-y
ISSN 1863-2386
ESSN 1863-2394

Publisher: Springer

The final publication is available at Springer via http://dx.doi.org/10.1007/s11761-019-00264-y

Improving generalization ability of Instance-transfer Based Imbalanced Sentiment Classification of Turn-Level Interactive Chinese Texts

Feng Tian [1] *, Prof; Fan Wu [2] *, Ms.; Xiang Fei [3], PhD; Nazaraf Shah [4], PhD; Qinghua Zheng [5], Prof; Yuanyuan Wang [6], Ms.

* These authors contributed equally to this work

[1] Systems Engineering Institute, Xi'an Jiaotong University, Xi'an, 710049, China.

E-mail: fengtian@mail.xjtu.edu.cn

[2] Systems Engineering Institute, Xi'an Jiaotong University, Xi'an, 710049, China.

E-mail: feeling_fan@163.com

[3] Faculty of Engineering and Computing, Coventry University, Priory Street, Coventry CV1 5FB, United Kingdom. E-mail: aa5861@coventry.ac.uk

[4] Faculty of Engineering and Computing, Coventry University, Priory Street, Coventry CV1 5FB, United Kingdom. E-mail: aa0699@coventry.ac.uk

[5] Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, 710049, China.

E-mail: qhzheng@mail.xjtu.edu.cn

[6] Systems Engineering Institute, Xi'an Jiaotong University, Xi'an, 710049, China.

E-mail: gothic@163.com

Corresponding author: Fan Wu, Systems Engineering Institute, Xi'an Jiaotong University, No 28, Xian Ning Xi Road, 710049, Xi'an, China.

E-mail: feeling_fan@163.com

Acknowledgments

This work is supported by National Key Research and Development Program of China (2018YFB1004500), National Nature Science Foundation of China (61877048614723l5), Innovative

Research Group of the National Natural Science Foundation of China (61721002), Innovation Research Team of Ministry of Education (IRT17R86), Project of China Knowledge Center for Engineering Science and Technology. Project of Chinese academy of engineering ”The Online and Offline Mixed Educational Service System for ’The Belt and Road’ Training in MOOC China”.

Abstract

Generally, a classification model achieving better generalization ability means the model performs better on the future-incoming data, otherwise the history dataset. Increasing the generalization ability of multi-domain and imbalanced multi-class emotion classification of turn-level interactive Chinese texts poses the challenges due to its high dimension and sparse feature values in its feature space. Moreover, the properties of different feature spaces or diverse data distributions in various domains of target dataset (T) and source dataset (S) make it difficult to employ multi-class and multi-domain instance transfer. To address these challenges, we propose a data level sampling approach for multi-class and multi-domain instance transfer which is inspired by transfer learning. To verify validity of our proposed method, an imbalanced dataset is taken as target dataset, while three datasets, one collected from Bulletin Board System of Xi'an Jiaotong University and other two datasets collected from China microblog platform Weibo, as source datasets. The experimental results show that the proposed approach outperforms classic algorithms by alleviating the imbalanced problem in interactive texts effectively. Moreover, a classification model that is trained on immigrated datasets produced by employing our proposed method achieves the best ability of generalization.

Key words: imbalanced sentiment classification, multi-class, multi-domain, interactive Chinese texts, instance immigration-based sampling, generalization ability

1. INTRODUCTION

Interactive text is an important form of communication on social media (such as Micro-blog comments, instant messaging, BBS post etc.) [1] [2]. During the conversation over internet, users express their views and emotions on the basis of turn-level text, emoji, pictures and so on [3]. Therefore, a strong time dependence characteristic of this kind communication leads to the issue of over-fitting to historical/trained data and a poor performance on new data in classification model. This problem is called for poor generalization ability of interactive text classification model. The issue compounded with short, incomplete and incoherent Chinese interactive texts results in the greater challenges in modeling.

In recent years, more and more researchers have paid attention to the topic and emotion recognition of interactive text [4] [5] [6] [7]. The existing research approaches to interactive texts rely on an assumption that the distribution of classes in each emotion recognition application is balanced. However, in realistic scenarios, the imbalanced data encountered in classification is a common problem, especially when the size of majority classes is above three times of the size of minority classes. This highlights the problem that the minority class information tends to be ignored during the training phase of classification model. This leads to the model trained from this kind of dataset having low identification precision in minority classes, which is also known as over-fitting for majority class. In our previous research, we have applied an in- stance transfer method to the emotion imbalanced product reviews. In which, a function is employed to choose features for evaluating the instance similarity between source and target datasets. The function calculates the sum of the information gains of Top-N common features of these two datasets and their proportions in the sum. Moreover, a homogenization processing method based on SMOTE is presented for feature spaces of the target dataset and the source dataset to overcome the feature spaces inconsistency between these two datasets. The proposed method effectively alleviates the imbalanced problem in target dataset [8]. However, previous research did not focus on the generalization ability of the classification model. The turn-level interactive texts have the characteristics of time dependence (i.e., topic and emotion change with time), class distribution imbalance, short sentences, lack of sentence constituents, richness of nonverbal signs [2] [7], which lead to the following difficulties of turn-level interactive text sentiment classification:

1. Compared with the datasets (product reviews) used in previous studies, the sentence length of interactive text is much shorter. For example, the sentence length of most interactive texts is less than 20 words and often lack of sentence constituents, which result in sparse feature values of high feature dimension. This make the existing homogenization processing method based on SMOTE [8] perform inefficiently.

2. As the topic and emotion change over time in interactive text, selecting suitable auxiliary dataset from mass data is of a vital important. Source datasets directly affect the characteristics of immigrated datasets and has great potential influence on the generalization ability of the trained model. At the same time, text datasets produced by different sources are mostly heterogeneous datasets (inconsistent feature space). How to evaluate the similarity of two heterogeneous datasets in order to select a suitable source dataset to be transferred is a big challenge. To our knowledge, there exist few preliminary researches on this topic.

Aiming at addressing the above difficulties, this paper proposes a multi-domain and multi-class instance immigration approach for imbalanced emotion classification of turn-level interactive Chinese texts, on the basis of the framework in the previous study [8]. The main contributions of this paper are as follows:

1. This paper proposes a new similarity measure method based on the sum of weighted KL

divergence of common features to select the suitable source datasets from multiple candidate datasets and measure the similarity of target and source datasets.

2. A new feature space homogenization method based on unique features of target dataset and the cosine similarity score of common features is proposed to overcome the sparse feature values of high feature dimension in interactive text.

3. Considering the feasibility of different common feature selection methods and instance similarity measurements, hundreds of comparative experiments on multi-domain and multi-class instance immigration are carried out. Through the experimental results, we select the best combination of instance selection methods.

Note that, the datasets we used in this paper contain two similar scale of minority emotion classes. The terms, sentiment and emotion are used interchangeably, so there is no difference between them in this paper [7].

2. RELATED WORK

This section presents the related work on different sentiment classification tasks, sentiment classification methods and imbalanced data classifications. According to the granularity of the processed texts, there are five levels of sentiment classification task: word- level, phrase-level, sentence-level, paragraph-level and document- level.

- Word-level, also called sentiment lexicon construction. In [9] a method which learns subjective nouns through semantic orientation of surrounding texts is proposed. In [10], the authors presented a method based on the pointwise mutual information of semantic orientation to infer the polarity of words according to the association with seven standard words. The authors in [11] used SHAL space to describe the polarity space of each word and improved the sentiment analysis on semantic orientation.

- Phrase-level. In [12] the phrase level sentiment analysis by adopting a two-phase classification method is explored, which first determines whether an expression is neutral or polar and then disambiguates the polarity of the polar ex- pressions. They also evaluated the performance of multiple features, including word features, modification features, sentence features, structure features and document features across multiple machine learning algorithms.
- Sentence-level. [13] focused on the subjectivity of sentences in close proximity to the sentence of interest, while other sentence-level methods [14] [15] [16] [17] analyzed the polarity of evaluating units, including words and sentences, and their combinations in a paragraph.
- Paragraph-level. In [13], the authors introduced machine learning approaches for the paragraph-level task. In [16], the Naive Bayes classifier is applied to classify the opinions in paragraphs.
- Document-level. Sentiment classification research in document level has been widely carried out [18] [19] [20] [21]. [19] compared the document-level sentiment analysis performance of NB-B (Naive Bayes using Bayes inference), NN-M (Naive Bayes using Maximum a posteriori) and SVM.

Existing research efforts on sentiment classification methods employ supervised machine learning techniques, such as Naive Bayes models, Decision Tree, Artificial Neutral Network and Sup- port Vector Machines (SVM). Recently, researchers have started to realize the importance of sentiment analysis on short texts or in sentence level. For example, the authors in [22] extracted sentiment strength from informal English text and used a method to exploit the de facto grammar and spelling styles of cyberspace. In [9], a Naive Bayes classifier using the subjective nouns is trained, discourse features, and subjectivity clues to distinguish the subjective sentences from objective sentences. In [23], the authors proposed a fine-to-coarse strategy for Chinese sentence- level sentiment classification based on sentiment dictionary. In [2], the authors had verified that three feature sets, syntactic feature set, frequency-based feature set and interaction-related feature set, help different classification methods to perform better in sentiment classification of turn-level interactive Chinese texts. However, it does not consider the issue of imbalanced classification. Moreover, imbalanced data classification is a challenging problem in the field of machine learning. The imbalanced distribution of class labeled samples (or class distribution) makes the classifier heavily biased towards majority class/label during the training process, which leads to a decrease in recognition performance [25]. Recently, the common methods to handle the above problem include data level sampling [24] [25] [26], cost sensitive learning [27] [28], feature selection [29] [30], feature weight adjustment [31] and one-class learning [32] [33] [34].

The research efforts mentioned above solve imbalanced problem aimed at a single target data set. It takes full use of the information of data itself to solve the problem. In recent years, with the development of transfer learning, researchers begin to adopt auxiliary datasets to solve the classification problem in different applications [35].

Since different text datasets are mostly heterogeneous datasets (the feature spaces of the two data sets are different), it is very important to solve the problem of how to measure the similarity of the datasets when selecting appropriate auxiliary dataset. In the field of data mining, the similarity measure of data samples / instances in the same feature space has been studied extensively [36], but the similarity measure of two heterogeneous datasets is seldom studied. The methods in [37] [38] are applicable to structured heterogeneous datasets, but the processing and feature extraction in these methods has a loss of semantic information and has a poor performance when conduct on text datasets.

3. MULTI-DOMAIN AND MULTI-CLASS INSTANCE IMMIGRATION

In multi-domain and imbalanced multi-class sentiment classification problem of turn-level interactive Chinese texts, the target dataset (T) could have different numbers of instances in different classes and domains. The number of instances in multi-class could have big difference among them (normally, less than 1:3~1:10) [39]. The core research idea of multi-class and multi-domain instance transfer is as follow:

According to the characteristics of the target dataset and classification target (currently, the main purposes of classification is to enhance the generalization ability of the classification model), select the suitable source dataset S from candidate datasets.

Considering that the multi-class classification task on datasets S and T is the same, we denote that the feature space in T and the one in S as Ω(F|T) and Ω(F|S) respectively, and then we transfer similar instances in S into T. In general, Ω(F|T)≠Ω(F|S). According to the existing research [2], the common features of interactive Chinese texts have syntactic, interactive and frequency feature. The most unique features are N-gram features. N-gram features refer to the combinations of the words and have a strong dependency on data/corpus. In this paper, Bigram is a subset of N-gram and adopted in syntactic feature set. The challenges to implementation of the core idea are how to evaluate the similarity and effectiveness of Ω(F|T) in T and Ω(F|S) in S, and how to overcome the inconsistent feature space between T and S which is caused by their unique features. It is imperative to solve following problems of: (1) select one suitable source dataset from the collected candidate datasets; (2) discovering and selecting common features of T and S; (3) evaluating the transfer ability of each instance in dataset S; (4) homogenizing incoherent feature spaces between transferred instances and dataset T to overcome issue of feature space inconsistency.

This paper proposes a new approach to solve the problems. The approach encompasses four steps:

Step 1: A similarity measurement method for heterogeneous datasets based on the sum of weighted KL divergence of common features is proposed. Calculate the sum of weighted KL of common feature; its value reflects the similarity degree of the candidate dataset and target dataset, the smaller the value is, the more similar they are. According to the classification target, improve the generalization ability of the model, and select the source dataset with bigger sum of weighted KL divergence of common features value, which can enrich the feature value. To improve the classification ability of the trained model for new data; to enhance the classification performance of the original dataset, the source dataset with smaller sum of weighted KL divergence of common features value.

Step 2: A greedy algorithm based on a function of calculating a proportion of sum of the information gain of Top- N common features between T and S is employed to solve the problem of discovering and selecting common features. Other indices for common feature selection such as term frequency-inverse document frequency (TF-IDF) [41] [40] [42] and Chi-square [43] have also been adopted for comparison;

Step 3: It evaluates the transferability of each instance in dataset S to determine appropriate instances to be transferred. It can be divided into two sub-problems: (1) Determining a suitable amount of the instances to be transferred; (2) Choosing appropriate instances from dataset S. To solve sub-problem (1), it starts with balancing the instance size of each class in difference domains of T to overcome their class imbalance. For the sub-problem (2), we adopt Cosine/Dice/Jaccard Index similarity scores based on common features to measure similarity between instances in S and the corresponding ones in T, while it needs to decide which instances in dataset S should be transferred to the corresponding domain in dataset T. This decreases the influence of instances to be transferred on the feature distribution of dataset T and increases the recognition precision;

Step 4: This step involves processing of the feature space inconsistency between the transferable instances from S and the ones in dataset T by combining the similar common features of T and S and feature space of T to solve the homogenization problem;

Step 5: It immigrates the transferable instances in S into dataset T by considering different domains and emotions in order to form a new target dataset D′ and it trains different classifiers on it and evaluate and com- pare their performances on the trained classification models to select the best one.

The following subsections describe the proposed method in details.

Section 3.1 describes the method of selecting suitable source dataset S. Section 3.2 describes the method of selecting the common features of both T and S. Section 3.3 presents a similarity calculation method for selecting the transferable instances from source dataset, which measures the transferability of each instance in S, while section 3.4 introduces the homogenization process for the feature space of transferable instances in S.

3.1. Similarity measurement method of datasets based on weighted KL divergence of common features

For a particular target dataset, according to the characteristics of its data source, selecting more suitable dataset for instance immigration from numerous auxiliary datasets is conducive to enhance the quality of transferred data, and lay a high-quality foundation for improving the classification performance of the model. In general, it is difficult to obtain the implicit features of a data source (including participants group category, social status, age, education etc.) which can well reflect the differences of datasets. However, by analyzing the feature distribution of the common features of the collected datasets, it can reason out that the diversity of the feature distribution reflects the overall diversity of datasets to some extent. Based on this finding, we propose a method to estimate the similarity of two heterogeneous datasets by computing the distance of the distribution of their common features. At present, there are mainly KL divergence, Bhattacharyya distance, Earth mover’s distance and so on for the similarity computation of distribution. In this paper, we mainly use KL divergence (Kullback-Leibler divergence). For discrete probability distributions P and Q, the KL divergence from Q to P is defined [44] to be:

$$D_{KL(P||Q)} = \sum_i P(i) log \frac{P(i)}{Q(i)} \qquad (1)$$

Where. p and q denote the densities of P and Q.

Considering that each feature has different importance in classifications, we should take it into account when select dataset. Therefore, the importance of features in classification is introduced into the similarity measure of datasets, and the KL divergence of each common feature distribution is weighted by feature importance. In this paper, the information gain of common features is used as weight value， noted $w_{inforgain}$. The sum of weighted KL divergence (SoWKL) of common features is noted as $sum_{weighted-kl}$. The final formula of the sum of the weighted KL divergence is as follows:

$$sum_{weighted-kl} = \sum_j w^j_{inforgain} D^j_{KL(P||Q)} \qquad (2)$$

The main steps are as follows:

- Estimate the probability distributions of the common features of the source dataset and the target dataset respectively. Common features include a variety of data types (numeric, nominal etc.). For different data types, the corresponding method to estimate the probability distribution of features is proposed. (1) For nominal data, the probability distribution is estimated directly through the frequency of each attribute value; (2) For numeric data, interval partitioning is applied to feature values. Since the object of this study is interactive text, the numerical features of the short text are mainly the frequency of certain grammatical structures and the frequency of some collocation. The analysis found that 90\% of the feature values are range between 0 and 10. So that, in combination with the length of the interactive text, we determine the segmentation points to be 1, 5, 10, and 20. After discretization, the distribution is calculated according to the nominal data processing.
- Calculate the KL divergence of common features of two datasets.

- Calculate the SoWKL of common features. Calculates the information gain of each feature in target dataset classification, and calculate the result according to formula 2.

3.2. Common features selection in source and target datasets

There are many indices for common feature selection. The research on sentiment classification of turn-level interactive Chinese texts, the research in [2][7] indicated that the method based on decision tree have good performance through ten-fold cross validation [45]. Information gain is a classical index for feature selection.

In our previous study [8], we have proposed a common feature selection method based on information gain. The steps of this method are as following:

- Compute the information gain of each feature in *T* and *S* respectively, and sort and list these features in descending order based on their information gain.
- Mark the position of common features in the sorted list.
- For each marked position, compute the proportion of the sum of information gain of a common feature located in the position in and the features lower than the position and the sum of information gain of all the features which appear before the position (that is called as the proportion of sum of the information gain of common features between T and S). Select the common features which have larger proportion to construct the features set to represent instances.

In this section, we adopt the previous method and steps in the recent study [8]. In addition, inspired by the feature selection indices of information retrieval, other indices include TF-IDF (term frequency-inverse document frequency [40][41][42]) and Chi-square [43] are also used for common feature selection.

Note that similar processes can be used when applying TF-IDF and Chi-square to be indices of common feature selection.

3.3. Selection of transferable instances from source dataset using similarity calculation rule

In the field of information retrieval, classical instance similarity calculations include cosine similarity, Dice Index and Jaccard Index [40]. In this section, we use cosine similarity as an example of selection methods for transferable instances from source dataset.

Cosine similarity is a common method for calculating two file similarity in natural language processing, in which each file is represented in a form of feature vector. This research adopts the cosine similarity scores based on common features to measure the similarity between instances in S and the corresponding ones in T, and to evaluate the transferability of instances in S. The algorithm can be divided into the following three steps:

Step 1 Express each instance with selected common features in a vector form and normalize them. The feature normalization process involves two sub steps: (1) Processing category attributes: Category attributes/features are replaced directly with numbers and the numerical value starting from 0 and increased by 1 subsequently. For example, the feature conjunction has 8 values: none, turn, casual, subjunctive, coordinate, comparison, undertake and conditional. We replace them with 0,1,2,3,4,5,6, and 7 respectively to convert the discrete quantities of the feature into numerical quantities; (2) Normalizing features: This adopts maximum and minimum normalization method [45] to normalize numerical features.

Step 2 Calculate the overall cosine similarity scores of corresponding emotion instances from the specific domain of source dataset and the emotion instances of the same domain in target dataset. Generally, the more similar two instances are, the higher their overall cosine similarity score is. Let $L = l_1, l_2, ..., l_N = l_p \mid p = 1, 2, ..., N$ denotes a set of class labels, *N* denotes the number of labels of classification tasks, $D = d_1, d_2, ..., d_M = d_k \mid k = 1, 2, ..., M$ denotes a set of domains (topics) in dataset, *M* denotes the number of the domains (topics), and the formula of cosine similarity calculation is as follows:

$$score(InsSou_{d_k}^{l_p}(\mathrm{i})) = \frac{\sum_{i=1}^{m} COS(InsSou_{d_k}^{l_p}(\mathrm{i}), InsTar_{d_k}^{l_p}(\mathrm{j}))}{m} \quad ......(3)$$

Where, $InsTar_{d_k}^{l_p}(\mathrm{j})$ denotes an instance labeled with l_p from a domain d_k in target dataset; $j = 1, 2, ..., n$ denotes that there are *n* instances with the same label in the same domain of the target dataset; $InsSou_{d_k}^{l_p}(\mathrm{i})$ denotes an instance labeled with l_p from a domain d_k in source dataset; $i = 1, 2, ..., m$ denotes that there are *m* instances with the same label in the same domain of the source dataset; $COS(InsSou_{d_k}^{l_p}(\mathrm{i}), InsTar_{d_k}^{l_p}(\mathrm{j}))$ means the common features-based cosine similarity score between $InsSou_{d_k}^{l_p}(\mathrm{i})$ and $InsTar_{d_k}^{l_p}(\mathrm{j})$, where the function $COS()$ calculates the cosine similarity between values of the common features of two instances after normalizing their feature values.

Step 3 The instances with same label from the same domains in source dataset are sorted by their cosine similarity scores based on common features in descending order, and the top ones have high priority for transfer.

Note that similar processes can be used when applying Dice and Jaccard to select transferable instances from source dataset.

3.4. Homogenization processing of feature space

Homogenization processing is used to solve the problem of incompatibility between the instances in source and target datasets. While the source and target datasets have common features, both T and S have unique features that lead to the situation where transferable instances from the source dataset cannot be used for training directly. Therefore, the homogenization processing should be carried out on the transferable instances to make the feature spaces of both T and S compatible. The elements and sizes of N-gram in T and S are different and their element types are all numerical. So the feature spaces of the immigrated instances can be unified by combining the cosine similarity *score* of N-gram based on common features in the corresponding domain. It ensures that the N-gram features of the transferred instances from source dataset and the N-gram features of the instances from the target dataset have a same dimension, while the common features can directly be used in new instances. The steps involved are as follows:

Step 1 As shown in Equation 4, calculate the average $\overline{NgT_{d_k}^{l_p}}$ of the features, N-gram, of each emotion class of different domains in target dataset respectively.

$$\sum Ng\_new(InsSou_{d_k}^{l_p}) = \overline{NgT_{d_k}^{l_p}} * score(NgS_{d_k}^{l_p}) + NgS_{d_k}^{l_p}(i^{'}) \text{ ......(4)}$$

Where $NgT_{d_k}^{l_p}$ denotes the value of the features, N-gram, of the *j*-th instance, which is labeled l_p from a domain d_k in the target dataset.

Step 2 As shown in Equation 5, construct new N-gram feature values of the transferred instances by combing their own values of the features N-gram with the average values of N-gram features in the target dataset, as well as their overall cosine similarity to make their feature space consistent with the target dataset.

$$Ng\_new(InsSou_{d_k}^{l_p}) = \overline{NgT_{d_k}^{l_p}} * score(NgS_{d_k}^{l_p}) + NgS_{d_k}^{l_p}(i)^{'} \text{ ......(5)}$$

Where, $NgS_{d_k}^{l_p}(i)$ denotes the value of the features N-gram of the i -th instance, which is labeled with l_p and belongs to d_k in the source dataset; $NgS_{d_k}^{l_p}(i)^{'}$ denotes filling the rest of the feature value by zero in order to keep the same dimensions of feature spaces of the transferable instances.

3.5. Instance combination and model training

The above three sections introduce how to select the instances to be transferred with the same label and from the corresponding domain of the source dataset and use the homogenization processing method to overcome the inconsistency of feature spaces between source and target datasets. Then, we transfer the instances selected from the source dataset into the target dataset to overcome the imbalanced problem in the target dataset. The next step is to train a sentiment classification model. The instance combination conforms to the following two principles:

- An instance from Domain S can only be transferred once, the reason is that multiple transfer of a same instance will cause over-fitting problem.
- It makes the number of instances balanced in each emotion class within the same domain in dataset T. That is to overcome the imbalance in each domain in the target dataset as much as possible.

4. EXPERIMENTS AND THEIR RESULT ANALYSIS

This section describes the steps involved in the experiments carried out and analysis of experimental results.

4.1. Experiment steps

Our experiments are described in the following steps:

Step 1: Collect corpora. We have collected five interactive Chinese text datasets in total. The name and instance number of each dataset is shown in Table 1. Linux_QQ and Linux_QQ_1030 are two chat log datasets of study groups, which were collected during different periods of time from an instant messaging tool named QQ. Xjtu_BBS is a posting record dataset collected from Bulletin Board System

(BBS) used in Xian Jiaotong University, China. Weibo1 and Weibo2 are two microblog datasets collected from China's biggest microblog platform, Weibo (weibo.com). These five datasets are denoted as Q, Qt, B, W1 and W2. After each turn in these corpora was labeled manually with emotion and domain categories, a statistical analysis of Q, B, W1 and W2 was carried out as shown in Figure 1. It can be observed from Figure 1 that datasets Q is an imbalanced corpus as imbalanced problem exists in each of their domains (topics); datasets B, W1 and W2 have rich domain (topic) knowledge, especially they have a large number of potential transferable instances that are targeted for minority classes in Q. We conduct the experiments in the following steps 2-7. In which, Q is taken as target dataset, while three datasets, including B, W1 and W2, act as source datasets. Each turn in these datasets was parsed and its features were abstracted by using approaches proposed in our prior work [2] [7]. Qt is the testing dataset to evaluate the generalization ability of each classification model.

Insert Table 1

Insert Figure 1 a-d

Step 2: Measure the similarity of target and source datasets according to steps mentioned in Section 3.1. Calculate the SoWKL of common features of Linux_QQ with Xjut_BBS, Weibo1, Weibo2.

Step 3: Select common features according to steps mentioned in Section 3.2 and calculate their overall similarity in each domain according to the steps described in Section 3.3, and then determine the instances to be transferred in each domain.

Step 4: Carry out the feature space homogenization processing method on the instances to be transferred according to the steps presented in Section 3.4.

Step 5: Incorporate the transferred instances into each domain of the target datasets according to steps described in Section 3.5 and form new training datasets by employing different indices of common feature selection and different instance similarity calculation of the transferability of instances. SMOTE is applied to each class of the new training datasets to make their class distribution balanced if the size of each class in the immigrated datasets is still not balanced. Note that the information gain [45], TF-IDF (term frequency- inverse document frequency [40], [41] [42]) and Chi- square [43] test are adopted to select common features and calculate cosine similarity as well as Dice Index and Jaccard Index are employed to compute the similarity of following format: name of target dataset name of source dataset name of common feature selection method name of similarity calculation method. To shorten the length of each name, information gain, TF-IDF and Chi-square are denoted as infg, tfidf and chi, while cosine similarity, Dice Index and Jaccard Index are denoted as cos, dice and jac. The immigrated datasets are represented as infg cos, infg dice, infg jac, tfidf cos, tfidf dice, tfidf jac, chi cos, chi dice and chi jac. For comparison with traditional data sampling strategies/methods for imbalanced datasets, Subsampling and SMOTE [46] [47] are produced respectively. Subsampling represents the dataset processed by subsampling method that select certain number of instances at most in each emotion class. SMOTE represents the dataset processed by the SMOTE method. All datasets associated with our experiment are listed in Table 2.

Insert Table 2

Step 6: Measure the similarity of target and source datasets according to steps mentioned in Section 3.1. Calculate the SoWKL of all features of Linux_QQ_1030 and every immigrated dataset.

Step 7: Evaluate the ability of generalization of the classification models. We take Linux_QQ and its immigrated datasets as training set for the classification method and Linux_QQ_1030 as a test dataset. five classical algorithms, Random Committee, Random Forest, Libsvm, Navie Bayes and J48 have been adopted in this step. In this step, we use ROC to evaluate the performance of the classification models.

4.2. Experiment results

The results of the experiments are described as follows.

In the experiment, the feature space of these four datasets is classified into three kind feature sets: syntax feature set, frequency- based feature set and interaction-related feature set [2]. The total number of features in Linux_QQ, Xjtu_BBS, Weibo1 and Weibo2 are 1751, 889, 1243 and 1200 respectively, and the number of common features of Linux_QQ and Xjtu_BBS is 38. The number of common features of Linux_QQ and Weibo1 is 33 and the number of common features of Linux_QQ and Weibo2 is 33. The selected features of each T and S group by using the algorithms described in Section 3.2 are shown in Table 3. The common features are selected according to the index of information gain. We also use Chi-square and TF-IDF as the comment feature selection method. In addition, we also adopted Dice index and Jaccard index as a measure of instance similarity calculation.

Insert Table 3

Calculate the SoWKL of Q with Linux_QQ_1030, Xjtu_BBS, Weibo1, and Weibo2 respectively. The results are shown in Table 4.

Insert Table 4

The number of the transferred instances from each domain (topic) in the source dataset is shown in Tables 5-7. As the size of each class in study domain of the immigrated dataset is still not balanced when set Xjtu_BBS is source dataset, SMOTE is applied to each emotion class of new dataset transferred from Xjtu_BBS.

Insert Table 5

Insert Table 6

Insert Table 7

The distribution of the instances in each domain and emotion of immigrated datasets is shown in Figures 2-4 according to Step 2. Calculate the SoWKL of Linux_QQ_1030 with every immigrated dataset, respectively. The results are shown in Table 8. Other experimental results are shown in Appendix Table 1. Table 9 show the experimental results corresponding to Step 7 of our experiments. In this step, we adopted five classification algorithms. After carrying out the experiments, we only list the weighted average ROC of several immigrated datasets which have made the classification model perform better on Qt. Other experimental results are shown in Appendix Table 2.

Insert Table 8

Insert Figure 2

Insert Figure 3

Insert Figure 4

Insert Table 9

4.3. Analysis of experimental results

The distribution of three emotion classes in each domain of immigrated datasets are shown in Figure 2-4. It is obvious that the immigrated datasets have the same number of instances in the three emotion classes. The proposed method can alleviate the imbalanced emotion distribution.

In order to compare the overall performance improvement when dealing with the results of a group of experiments transferring different source datasets into a target dataset, we de- fine an index called average performance improvement (AvPI). A percentage of AvPI (PAvPI) is equal to the average of the difference of five method's performance on a dataset (such as each immigrated dataset, SMOTE dataset, subsampling dataset) dividing by performance on Q. To evaluate the performance of the proposed method on solving the imbalance class problem, we analyzed the experimental results of the weighted average of ROC.

Table 7 and Appendix Table 1 show the generalization experiment's results of five classification algorithms (Naive Bayes, Random Committee, Random Forest, LibSVM and J48) on 30 datasets related to Linux_QQ when taking Linux_QQ_1030 as a test dataset. Compared with the weighted average of ROC in Linux_QQ, PAvPI of Q_Subsampling is 0.04% (that means that it has an average of 0.04% increase on Q_Subsampling) and PAvPI of Q_SMOTE is -2.8%, while the values of PAvPI in immigrated dataset Q_B_Chi_cos, Q B inforgain cos, Q_B_tfidf cos are 5.96%, 5.55% and 4.71%, respectively. The best performance of the classification results is achieved when conducted Nave Bayes on immigrated dataset Q_B_inforgain_cos. Therefore, it can be concluded from the experimental results that the proposed method is effective and superior to SMOTE and subsampling.

The results in Table 9 also show that the datasets constructed by Q and B have greater values of PAvPI. Majority of datasets, which are immigrated from Xjtu_BBS, help the five classification algorithms to outperform the 18 datasets immigrated from Weibo1 and Weibo2. Furthermore, it can observe in Table 4 that Xjtu_BBS is the most similar one to the target dataset Linux_QQ among three candidate datasets, as well as in Table 8 that the immigrated datasets constructed by Xjtu_BBS are more similar to test dataset Linux Q 1030 among 27 candidate datasets (seen in Appendix). Therefore, we can conclude that, for the target dataset Linux_QQ, the optimal source dataset is Xjtu_BBS. This conclusion also has been verified according to the results of Table 9. Meanwhile, according to in Tables 8 and 9, the experiment results show that the proposed common feature weighted KL can measure the similarity of two datasets and give a clear clue to select the suitable candidate dataset as a source dataset in our experiments.

According to ascending order of SoWKL, we draw the SoWKL curve of 27 immigrated datasets and the corresponding AvPI curve in the same figure, as shown in Figure 5. It is obvious that when SoWKL value is less than 0.013, the corresponding AvPI is positive which means the classification model trained on that immigrated datasets can enhance the generalization ability, and when the SoWKL value is greater than 0.016, it will reduce the generalization ability of classification model. Moreover, compared with other classification algorithms, Naive Bayes used on immigrated datasets classification can achieve better performance. By analyzing the experimental data, we find that the data types of experimental data are mainly discrete, and Naive Bayes has good classification effect for discrete data. This makes the overall classification performance of immigrated datasets on Naive Bayes better than that of other classification algorithms.

Insert Figure 5

Based on the experimental results, we can conclude that: ***a.*** The proposed method can promote the generalization ability of multi-domain and imbalanced multi-class imbalanced sentiment classification of turn-level interactive Chinese texts. ***b.*** Source dataset has an obvious influence on the performance of classification model. By computing the SoWKL of common features between target and candidate datasets, we can select the most suitable source dataset. In addition, when the SoWKL of immigrated dataset and Qt is less than 0.013, the generalization ability of the classification model can be improved efficiently. ***c.*** According to our experimental results, combining information gain, cosine similarity and Naive Bayes can achieve the best classification performance.

5. CONCLUSIONS

Interactive text is an important target object of sentiment classification. The characteristic of its imbalanced class distribution poses many challenges to turn-level sentiment classification. Moreover, increasing the generalization ability of the classification model to achieve better performance on the future-incoming data is vital important. This paper attempts to address these challenges by proposing multi-class and multi-domain instance immigration approach for imbalanced sentiment classification of turn-level interactive Chinese text. The main contributions of this paper are as follows:

Firstly, a similarity measurement method for heterogeneous datasets based on the sum of weighted KL divergence of common features is proposed to select the most suitable source dataset. Secondly, a greedy algorithm based on a function of calculating a proportion of sum of the information gain of Top-N common features between T and S is employed to solve the problem of discovering and selecting common features. Thirdly, a method to evaluate the transferability of each instance based on the similarity calculation of the common features of transferable instances in the same domain of source and target datasets is used. It can solve two sub-problems: (1) Determining a suitable amount of the instances to be transferred; (2) Choosing appropriate instances from dataset S. To solve sub-problem (1), it starts with balancing the instance size of each class in difference domains of T to overcome their class imbalance. Finally, in order to solve the feature space inconsistency between the transferable instances from S and the ones in dataset T, a homogenization processing is proposed.

The experimental results clearly indicate that our approach provides a better solution for multi-domain and multi-class in- stance immigration when the feature space of target dataset and source dataset is non-homogeneous, and can improve the generalization ability of classification model. And combine information gain, cosine similarity and Naive Bayes can achieve the best classification performance. Its performance is superior to some classical methods such as SMOTE and subsampling.

In addition, we find that source dataset has an obviously influence on the performance of classification model. By calculating the SoWKL of candidate datasets and target dataset, and selecting appropriate source dataset, the generalization ability of the classification model can be adjusted in a quantitative manner.

Based on the research results of this paper, the future work will aim at proposing more comprehensive datasets similarity measurement parameters with multi-source instance immigration. Through the feedback adjustment of parameters, select more suitable instances from multiple source datasets and optimize the multi-class and multi-domain instance immigration algorithm, and finally make further efforts on improving the generalization ability of classification model.

Compliance with Ethical Standards

Funding:

This study was funded by National Key Research and Development Program of China (6187704861472315), Innovative Research Group of the National Natural Science Foundation of China(61721002), National Nature Science Foundation of China(61877048，61472315), Innovation Research Team of Ministry of Education(IRT_17R86).

Conflicts of Interest:

The authors declare that they have no conflict of interest.

REFERENCES

1. Tian F., and Zheng Q., and Zheng D.,Mining Patterns of e-Learner Emotion communication in Turn Level of Chinese Interactive Text Experi- ments and Findings,Proceeding of 2010 12th International Conference on Computer Supported Cooperative Work in Design. Shanghai, China:664- 670, 2010.
2. Tian F. and Liang H., and Li L.,Sentiment Classification in Turn-Level Interactive Chinese Texts of E-learning Applications,Proceeding of 2012 12th International Conference on Computer Supported Cooperative Work in Design. Roma, Itali:480-484, 2012.
3. Gibson W., Intercultural Communication Online Conversation Analysis and the Investigation of Asynchronous Written Discourse, Forum Qualita- tive Social Research. 10(1):1-18, 2009.
4. Wu C., Huang Y. and Hwang J., Review of affective computing in education/learning: Trends and challenges, British Journal of Educational Technology. 47(6):13041323, 2016.
5. Liu Z., Liu S., Liu L.,Sun J., Peng X., Wang T., Sentiment recognition of online course reviews using multi-swarm optimization-based selected features, Neurocomputing. 185:1120, 2016.
6. LiuZ.,ZhangW.,SunJ.,HercyN.H.Cheng,PengX.,LiuS.,Emotionand Associated Topic Detection for Course Comments in a MOOC Platform, 2016 Interna-tional Conference on Educational Innovation through Tech. 2016.
7. Tian F., Gao P, Li L., Zhang W, Liang H., Qian Y., Zhao R., Recognizing and regulating e-learners emotions based on interactive Chinese texts in e-learning systems, Knowledge-Based Systems. 55:148-164, 2014.
8. Tian F., Wu F., Chao K, Zheng Q, Shah N., Lan T., Yue J., A Topic Sentence-based Instance Transfer Method For Imbalanced Sentiment Clas- sification of Chinese Product Reviews, Electronic Commerce Research and Applications. Vol16:6676, 2016.
9. Riloff E., Wiebe J. Wilson, T.,Learning subjective nouns using extraction pattern bootstrapping, Proceedings of the Seventh Conference on Natural Language Learning Conference. Edmonton, Canada:25-32, 2003.
10. Turney P., Littman M., Measuring Praise and Criticism Inference of semantic orientation from association, ACM Transactions on Information Systems. 21(4):315-346, 2003.
11. Xu T., Peng Q., Identifying the semantic orientation of terms using S- HAL for sentiment analysis, Knowledge-based systems. 35:279-289, 2012.
12. Wilson T., Wiebe, J., Hoffmann, P., Recognizing Contextual Polarity An Exploration of Features for Phrase-level Sentiment Analysis, Computa- tional Linguistics. 35(3):399-433, 2009.
13. Pang B., Lee L., A sentimental education: Sentimental analysis using subjectivity summarization based on minimum cuts, Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain:271-278, 2004.
14. Kim S., Hovy, E., Automatic detection of opinion bearing words and sentences, Proceeding of the IJCNLP 2005. Jeju Island,Republic of Korea:61-65, 2005.
15. Yin C., Peng Q.,Sentiment Analysis for Product Features in Chinese Reviews Based on Semantic Association, Proceeding of International Con- ference on Artificial Intelligence and Computational Intelligence 2009. Shanghai, China:81-85, 2009.
16. Hatzivassiloglou V., Wiebe J., Effects of adjective orientation and gradability on sentence subjectivity, Proceeding of the International Conference on Computational Linguistics

(COLING). Sarbrucken, Germany:299-305, 2000.

17. Yu H., Hatzivassiloglou V., Towards answering opinion questions separating facts from opinions and identifying the polarity of opinion sentences, Proceeding of the EMNLP 2003. Sapporo, Japan:129-136, 2003.

18. Efron M., Cultural orientations Classifying subjective documents by cocitation analysis, Proceedings of the AAAI Fall Symposium Series on Style and Meaning in Language, Art, Music, and Design. Washington DC, USA:41-48, 2004.

19. Lin W., Wilson T., Wiebe J., Which side are you on? Identifying perspec- tives at the document and sentence levels, Proceeding of the Conference on Natural Language Learning. Morristown, USA:109-116, 2006.

20. Jindal N., Liu B., Identifying comparative sentences in text documents, Proc. of the ACM Special Interest Group on Information Retrieval. Seattle, USA:244-251, 2006.

21. Khan K., Baharudin B., Khan A., Malik, F., Mining Opinion from Text Documents A Survey, Proceedings of 2009 3rd IEEE International Confer- ence on Digital Ecosystems and Technologies. Istanbul, Turkey:217-222, 2009.

22. Thelwall M., Buckley K., Paltoglou G., Cai D., Kappas, A., Sentiment Strength Detection in Short Informal Text, Journal of the American Society for Information Science and Technology. 61(12):2544-2558, 2010.

23. Wang X., Fu G., Chinese Sentence-Level Sentiment Classification Based on Sentiment Morphemes, 2010 International Conference on Asian Lan- guage Processing. Harbin, China:28-30, 2010.

24. Borrajo L., Romero R., Iglesias E., Redondo, E., Improving imbalanced scientific text classification using sampling strategies and dictionaries, Journal of Integrative Bioinformatics. 8(3):176-191, 2011.

25. Barandela R., Valdovinos R., Sa´nchez J., Ferri F., The imbalanced training sample problem Under or over sampling?, Berlin,Heidelberg: Springer. Lecture notes on computer science volume 3138 Structural, Syntactic, and Statistical Pattern Recognition, 2004.

26. Chawla N., C4.5 and imbalanced data sets investigating the effect of sampling method, probabilistic estimate, and decision tree structure, Pro- ceedings of the ICML03, Workshop in Datasets Imbalance. Washington DC, USA:315-330, 2003.

27. Kamel M., Wong A., Wang Y., Cost sensitive boosting for classification of imbalanced data, Pattern Recognition. 40(12):3358-3378, 2007.

28. Zhou Z., Liu X., Training cost-sensitive neural networks with methods addressing the class imbalance problem, IEEE Transactions on Knowledge and Data Engineering. 18(1):63-77, 2006.

29. WangS.,LiD.,SongX.,WeiY.,Li,H.,Afeatureselectionmethodbased on improved fishers´ discriminant ratio for text sentiment classification, Expert Systems with Applications. 38(3):8696-8702, 2011.

30. WangS.,LiD.,ZhaoL.,ZhangJ.,Samplecuttingmethodforimbalanced text sentiment classification based on BRC, Knowledge-Based Systems.37:451-461, 2013.

31. Liu T., Peng Q., Imbalanced text classification A term weighting approach, Knowledge based systems. 36(1):690-701, 2009.

32. Raskutti B., Kowalczyk A., Extreme rebalancing for SVMs a case study, ACM Sigkdd Explorations Newsletter. 6(1):60-69, 2004.

33. OguraH.,AmanoH.,KondoM.,Comparisonofmetricsforfeatureselec- tion in imbalanced text

classification, Expert Systems with Applications. 38(5):4978-4989, 2010.

34. SatyamM.,JitendraA.,SanjeevS.,ANewapproachforClassificationof Highly Imbalanced Datasets using Evolutionary Algorithms, International Journal of Scientific and Engineering Research. 2(7):1-5, 2011.

35. Pan S., Yang Q., A Survey on Transfer Learning, IEEE TRANSAC- TIONS ON KNOWLEDGE AND DATA ENGINEERING. 22(10):1345-1359, 2009.

36. Cha, Sung-Hyuk. Comprehensive survey on distance/similarity measures between probability density functions,City 1.2 (1), 2007.

37. Tatti N., Distances between data sets based on summary statistics, Journal of Machine Learning Research. 8(1):131-154, 2007.

38. Song Q., Wang G., Wang C., Automatic recommendation of classifica- tion algorithms based on data set characteristics, Pattern Recognition. 45(7):2672-2689, 2012.

39. He H., MA Y., Imbalanced Learning-Foundations, Algorithms, and Applications, IEEE Press. 2010.

40. Salton G., McGill M., Introduction to modern information retrieval, McGraw-Hill. 1983.

41. SaltonG.,FoxE.,WuH.,ExtendedBooleaninformationretrieval,ACM Communication. 26:1022-1036, 1983.

42. Salton G., Buckley C., Term-weighting approaches in automatic text retrieval, Information Processing and Management. 24(5):513-523, 1988.

43. Fisher R., Yates F., Statistical Tables for Biological, Agricultural and Medical Research, 6th Edition, Oliver & Boyd. 1963.

44. MacKay, David J.C., Information Theory, Inference, and Learning Algorithms (First ed.), Cambridge University Press. p34, 2003.

45. Han J., Kamber M., Data Mining Concept and Techniques, The Morgan Kaufmann. 2th Edition, 2006.

46. Chawla N., Bowyer K., Hall L., and Kegelmeyer W., SMOTE Synthetic Minority Oversampling Technique, Journal of Artificial Intelligence Research. 16:321-357, 2002.

47. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten, I.,The WEKA Data Mining Software An Update, SIGKDD Explorations. 11(1):10-18, 2009.

Tables

Table 1. Amount, simplified name and categories of collected datasets

Name	Simplified Name	The number of instances	Function
Linux_QQ	Q	5123	Target
Xjtu_BBS	B	9957	Source
Weibo1	W1	8417	Source
Weibo2	W2	8697	Source
Linux_QQ_1030	Qt	1030	Testing

Table 2. Datasets associated with our experiment

Method	subsampling	SMOTE	immigrated data
	Q_subsampling	Q_SMOTE	
B			Q_B_chi_cos
			Q_B_chi_dice
			Q_B_chi_jac
			Q_B_infg_cos
			Q_B_infg_dice
			Q_B_infg_jac
			Q_B_tfidf_cos
			Q_B_tfidf_dice
			Q_B_tfidf_jac
W1			Q_W1_chi_cos
			Q_W1_chi_dice
			Q_W1_chi_jac
			Q_W1_infg_cos
			Q_W1_infg_dice
			Q_W1_infg_jac
			Q_W1_tfidf_cos
			Q_W1_tfidf_dice
			Q_W1_tfidf_jac
W2			Q_W2_chi_cos
			Q_W2_chi_dice
			Q_W2_chi_jac
			Q_W2_infg_cos
			Q_W2_infg_dice
			Q_W2_infg_jac
			Q_W2_tfidf_cos
			Q_W2_tfidf_dice
			Q_W2_tfidf_jac

Table 3. Selected common features according to the index of information gain

Source datasets	Number	$F_{ST}^{'com}$
Xjtu_BBS	18	length,negFre, posFre, emotionGraph, mimeticExist, maxFre, nxExist, verbFre, adjBelongcomplement, punFre, advFre, emotionVerb, advBelongAdver, maxFre, conjunction, adjBelongAtt, adjBelongAdver, interjectionExist
Weibo1	17	oneFre,negFre, emotionVerb,FrecharFre, otherSign,verbFre, maxFre,adjBelongAtt, conjunction, advFre, function, twoFre, posFre, adjBelongAdver, advBelongAdver, adjBelongcomplement, nagatorBelongComplement
Weibo2	16	oneFre, negFre, emotionVerb,FrecharFre, otherSign, verbFre, maxFre, conjunction, posFre, twoFre, adjBelongAdver, topic, advBelongAdver, function, adjBelongcomplement, adjBelongAtt, nagatorBelongComplement

Table 4. the SoWKL value of Linux QQ with Xjtu BBS, Weibo1, and Weibo2

Candidate dataset	Xjtu_BBS	Weibo1	Weibo1
SoWKL	0.019	0.946	0.865

Table 5. The number of transfer instances from Xjtu_BBS to Linux_QQ according to cosine similarity

Topic	negative	positive	calm
Life	485	432	0
Study	820	1285	0
Love	19	16	0
Friend	17	11	0

Table 6. The number of transferred instances from Weibo1 to Linux_QQ according to cosine similarity

Topic	negative	positive	calm
Life	485	432	0
Study	3131	2919	0
Love	19	16	0
Friend	17	11	0

Table 7. The number of transferred instances from Weibo2 to Linux_QQ according to cosine similarity

Topic	negative	positive	calm
Life	485	432	0
Study	3131	2919	0
Love	19	16	0
Friend	17	11	0

Table 8. the SoWKL value of Linux_QQ_1030 with immigrated datasets

Candidate dataset	B_chi_cos	B_infg_cos	B_tfidf_cos
SoWKL	0.012267	0.00874	0.012039

Table 9. The experimental results corresponding to Step 7, the weighted average of ROC of five classification algorithms

Weighted Ave.	Random Committee	SVM	J48	Random Forest	NaIve Bayes
Q	0.675	0.619	0.608	0.672	0.706
Q_SMOTE	0.672	0.64	0.655	0.628	0.59
Q_Subsampling	0.651	0.622	0.682	0.609	0.704
Q_B_chi_cos	0.717	0.65	0.688	0.737	0.746
Q_B_infg_cos	0.717	0.647	0.655	0.738	0.75
Q_B_tfidf_cos	0.719	0.654	0.648	0.73	0.745

Figure legends

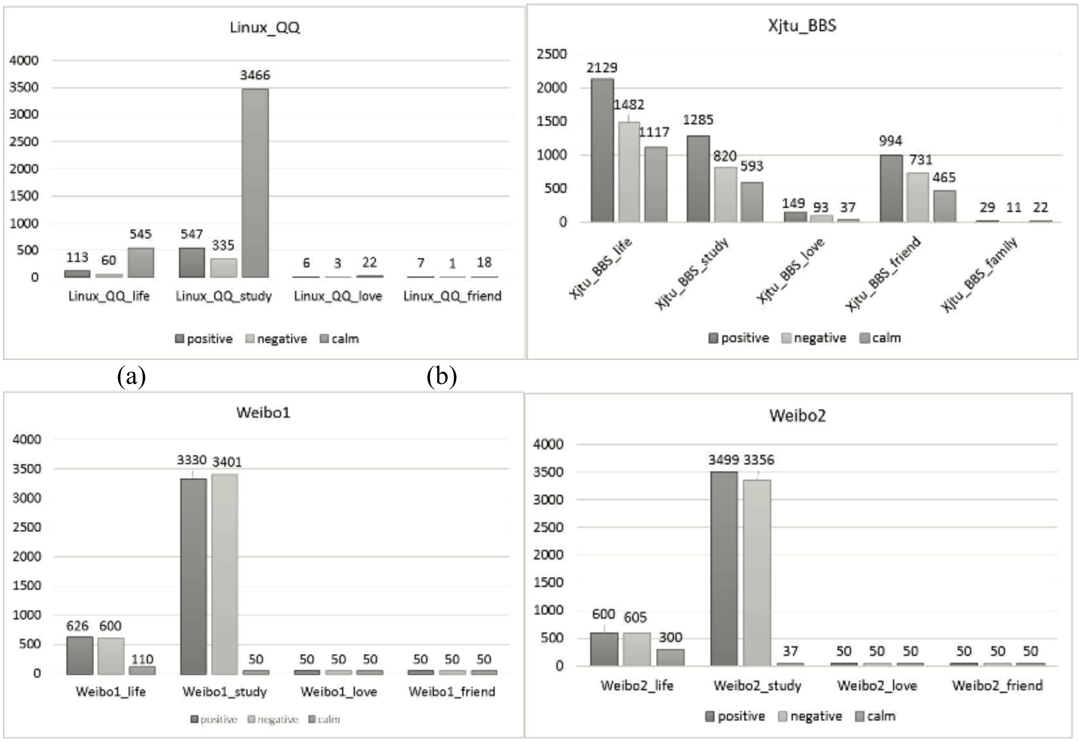
Fig 1: Instance distribution in domain and emotion of four datasets.

Fig 2: Emotion distribution of immigrated datasets of Linux_QQ and Xjtu_BBS.

Fig 3: Emotion distribution of immigrated datasets of Linux_QQ and Weibo1.

Fig 4: Emotion distribution of immigrated datasets of Linux_QQ and Weibo2.

Fig 5: SoWKL curve of 27 immigrated datasets and the corresponding AvPI curve.



(a) (b)

(c) (d)

Fig 1

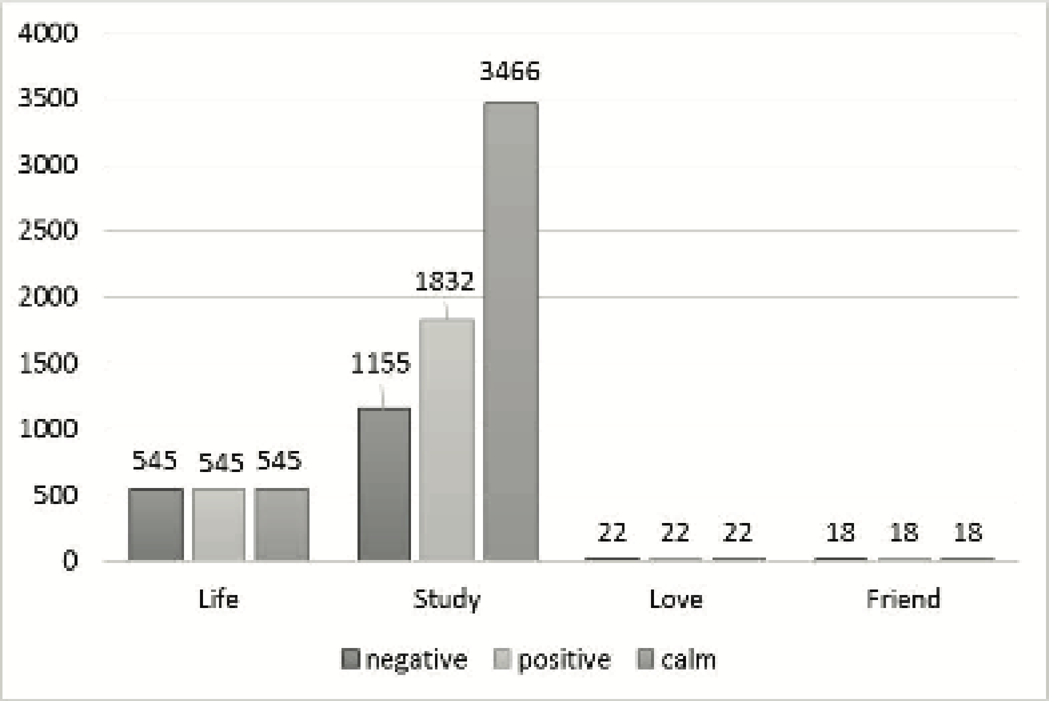


Fig 2

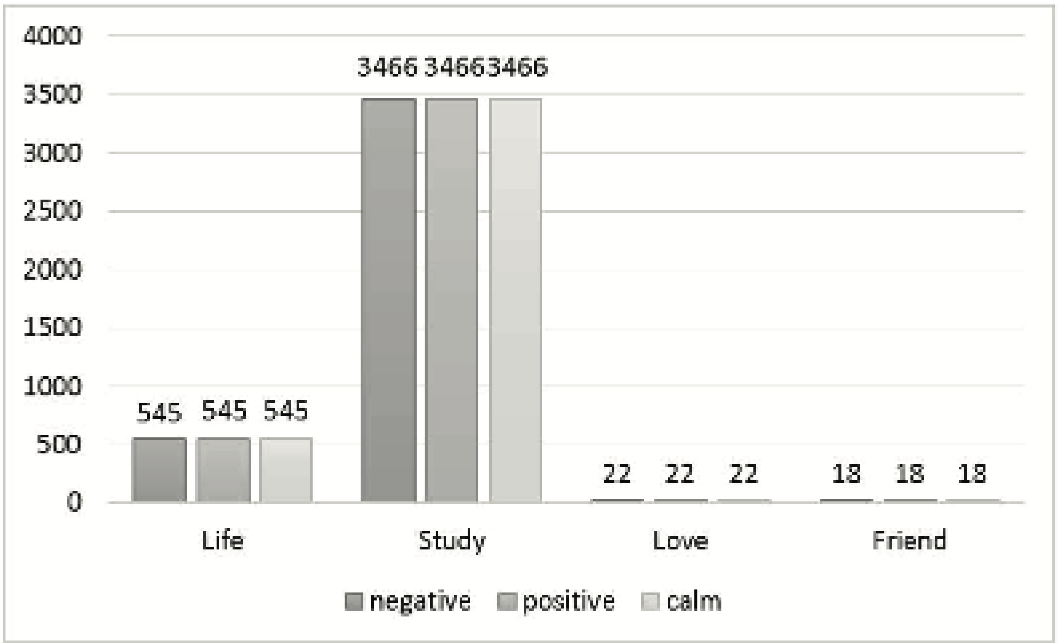

4000
3500
3000
2500
2000
1500
1000
500
0
545 545 545
3466 3466 3466
22 22 22
18 18 18
Life
Study
Love
Friend
negative
positive
calm

Fig 3

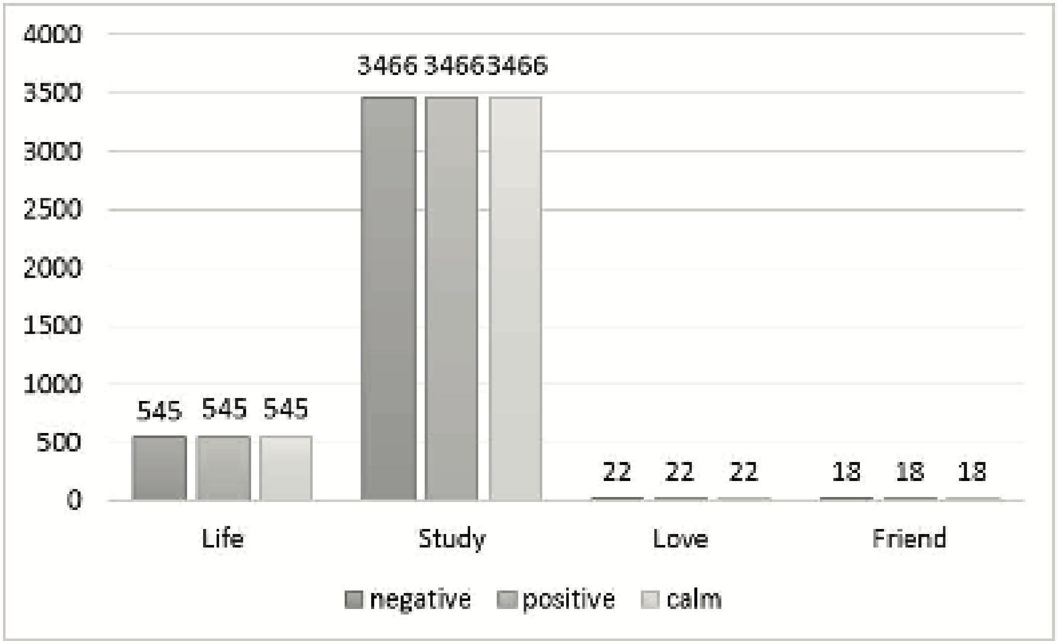

4000
3500
3000
2500
2000
1500
1000
500
0
545 545 545
3466 3466 3466
22 22 22
18 18 18
Life
Study
Love
Friend
negative
positive
calm

Fig 4

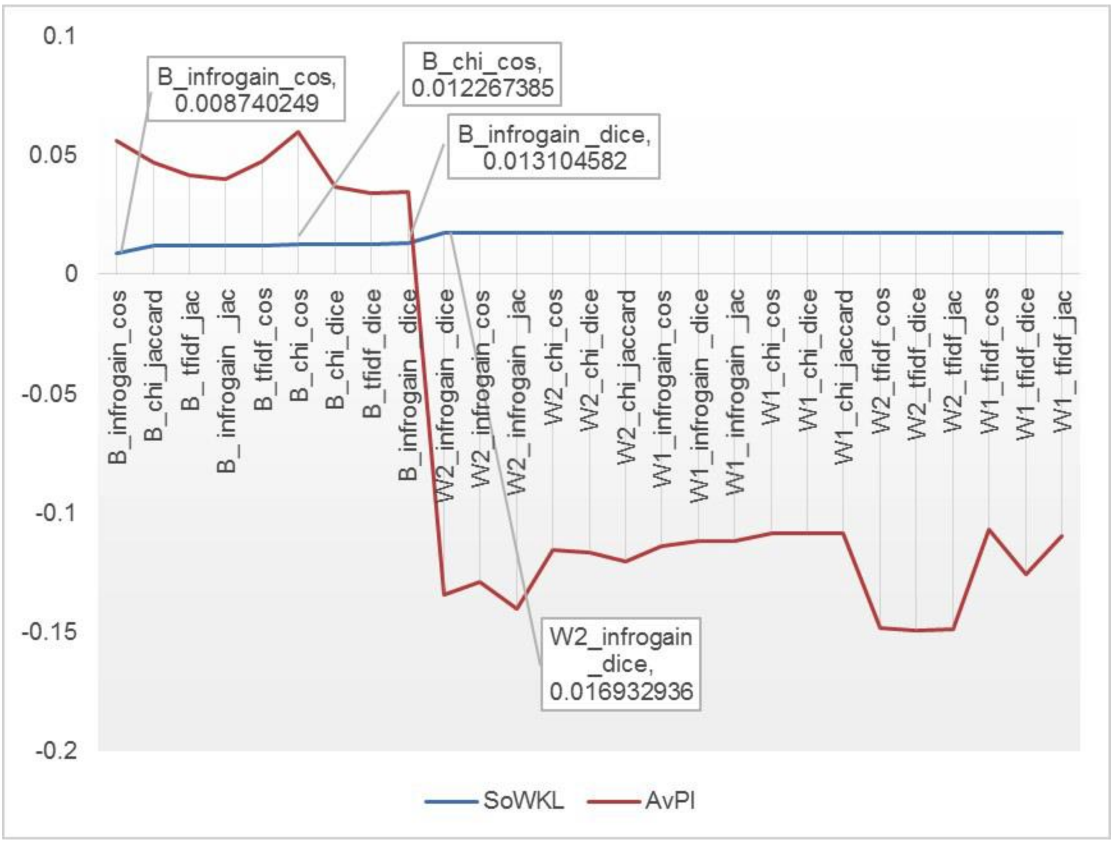

0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
B_infrogain_cos, 0.008740249
B_chi_cos, 0.012267385
B_infrogain _dice, 0.013104582
W2_infrogain _dice, 0.016932936
B_infrogain_cos
B_chi_jaccard
B_tfidf_jac
B_infrogain _jac
B_tfidf_cos
B_chi_cos
B_chi_dice
B_tfidf_dice
B_infrogain _dice
W2_infrogain _dice
W2_infrogain_cos
W2_infrogain _jac
W2_chi_cos
W2_chi_dice
W2_chi_jaccard
W1_infrogain_cos
W1_infrogain _dice
W1_infrogain _jac
W1_chi_cos
W1_chi_dice
W1_chi_jaccard
W2_tfidf_cos
W2_tfidf_dice
W2_tfidf_jac
W1_tfidf_cos
W1_tfidf_dice
W1_tfidf_jac
SoWKL
AvPI

Fig 5