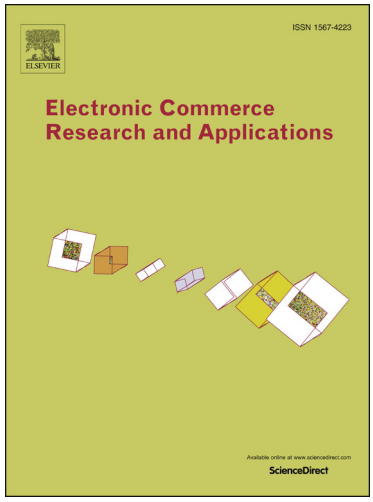


A Topic Sentence-Based Instance Transfer Method for Imbalanced Sentiment Classification of Chinese Product Reviews

Feng Tian, Fan Wu, Kuo-Ming Chao, Qinghua Zheng, Nazaraf Shah, Tian Lan, Jia Yue

PII: S1567-4223(15)00075-7
DOI: http://dx.doi.org/10.1016/j.elerap.2015.10.003
Reference: ELERAP 634

To appear in: *Electronic Commerce Research and Applications*

Received Date: 28 February 2015
Revised Date: 22 October 2015
Accepted Date: 22 October 2015

Please cite this article as: F. Tian, F. Wu, K-M. Chao, Q. Zheng, N. Shah, T. Lan, J. Yue, A Topic Sentence-Based Instance Transfer Method for Imbalanced Sentiment Classification of Chinese Product Reviews, *Electronic Commerce Research and Applications* (2015), doi: http://dx.doi.org/10.1016/j.elerap.2015.10.003

A TOPIC SENTENCE-BASED INSTANCE TRANSFER METHOD FOR IMBALANCED SENTIMENT CLASSIFICATION OF CHINESE PRODUCT REVIEWS

Feng Tian [a,b], Fan Wu [a,b], Kuo-Ming Chao [c], Qinghua Zheng [d],
Nazaraf Shah [c], Tian Lan [a,b], Jia Yue [a,b]

[a] Systems Engineering Institute, Xi'an Jiaotong University, Xi'an, China;
[b] Shaanxi Key Lab of Satellite-Terrestrial Network Tech. R&D, Xi'an Jiaotong Univ., Xi'an, China;
[c] Department of Computer Science and Technology, Coventry University, CV1 2JH, UK;
[d] Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China

Last revised: October 22, 2015

ABSTRACT

The increasing interest in sentiment classification of product reviews is due to its potential application for improving e-commerce services and quality of the products. However, in realistic e-commerce environments, the review-related data are imbalanced, and this leads to a problem in which minority class information tends to be ignored during the training phase of a classification model. To address this problem, we propose a *topic sentence-based instance transfer method to* process imbalanced Chinese product reviews by using an auxiliary dataset (source dataset). The proposed method incorporates a rule and supervised learning hybrid approach to identify a topic sentence of each product review and adds the feature set of the topic sentence to the feature space of sentiment classification. Next, to measure the transferability of instances in source dataset, a greedy algorithm based on information gain of top-N common features is used to select common features. Then, a common feature-based cosine similarity of instances between source dataset and target dataset is introduced to select the transferable instances. Furthermore, a *synthetic minority over-sampling technique* (Smote) based method is adopted to overcome feature space inconsistency between the source dataset and target dataset. Finally, we immigrate the instances selected in source dataset into target dataset to form a new dataset for the training of classification model. Two datasets collected from Jingdong and Dangdang are the target dataset and source dataset. The experimental results verify that, considering the ability of generalization, our proposed method helps a support vector machine (SVM) to outperform other classification methods, such as the J48, Naive Bayes, Random Forest and Random Committee methods, when applied to datasets produced by resampling and Smote.

Keywords: Classification methods, imbalanced sample classification, instance transfer methods, product reviews, topic sentence analysis

---

Acknowledgments. This research was partially supported by the National Natural Science Foundation of China under Grant Nos. 91118005, 91218301, 61103239, 61221063, 61428206 61472317 and 61532004, the Ministry of Education Innovation Research Team No. IRT13035, the National Key Technologies R&D Program of China under Grant No. 2013BAK09B01, the National High-Tech R&D Program of China (863 Program) No. 2012AA011003, Cheung Kong Scholar's Program, China Scholarship Council under Grant No. [2013] 3018 and the Innovation Project of Shaanxi Province Key Lab.

1. INTRODUCTION

In the last few years, we have witnessed a surge of interest in opinions mining automated systems for online product reviews. There are many representative articles in the large research literature (e.g., Bagheri et al. 2013, Fu et al. 2013, Zhang et al. 2012, Zhang et al. 2014). The major supporting technology for opinion mining systems includes topic modelling and sentiment analysis. Researchers, engineers, and practitioners believe that the systems capable of automatically analyzing consumer sentiment expressed widely in online venues will help companies to understand how the consumers perceive their products and services. Many research efforts on sentiment analysis on product reviews have been carried out to enable companies to understand consumer's perception of the products and services.

Most of them rely on an assumption that the class distribution in the training datasets is balanced. However, in reality, the class distribution in collected product review data is usually imbalanced, so they called *imbalanced data*. The imbalanced data encountered in classification is a well-known problem, especially when the size of majority classes is above three times of the size of minority classes. This leads to a situation where minority class information gets ignored during the training phase of a classification model. A model trained from this kind of dataset that have low identification precision in minority classes exhibit over-fitting of the majority class.

Some researchers have employed a sub-sampling strategy on imbalanced data to balance the class distribution of the dataset. This approach worsens the performance and generalization ability of the classification model trained on subsampled dataset. At the same time, different products (or topics, to relate products to the more formal language of our research) from one data source may have an imbalanced distribution in emotion classes. This may form different feature spaces with diverse data distributions in emotion classifi-

cation. That is, the imbalanced distribution of emotion classes with different topics represents different kind of interactions and mental states of the users.

The traditional methods for handling imbalanced classification problem rely on data level sampling, cost sensitive learning, features selection, feature weight adjustment and one-class learning approaches (Ogura et al. 2011). However, because these methods normally only rely on one dataset, the classification models that are trained on them have an over-fitting problem and lack the ability for generalization. For example, suppose that a balanced dataset is created from only one dataset according to a sampling strategy for training the classifiers. When a trained classifier is applied to a different real-world dataset for analysis, the classification performance is often degraded (He and Ma 2013).

Methodologies behind the classifiers that are trained on more than one auxiliary dataset have been widely adopted (Nguyen et al. 2011, Tommasi and Tuytelaars 2014, Gong et al. 2012, Heim et al. 2014, Hung and Lin 2011) in recent years in an attempt to address problems with insufficient and homogeneous data by adopting the knowledge transfer learning method (Pan and Yang 2010). A simple method may directly combine an auxiliary dataset and an original dataset into a single dataset to train the classifier. As the tasks of emotion detection are strongly domain and product- or topic-dependent. The feature distribution of each product will have its own characteristics. So we believe that such a method will destroy the innate and unique features that exist in different domains and will decrease recognition accuracy.

We are taking on the task of topic sentence-based instance transfer in this research. Our approach is to sample similar instances from the auxiliary dataset in order to deal with imbalanced sentiment classification of target dataset of product reviews. This can be classified as one of data level sampling approaches.

Figure 1 illustrates the core idea of this research on instance transfer for providing a solution to the problem of imbalanced sentiment classification of product reviews.

INSERT FIGURE 1 ABOUT HERE

We begin by defining some key language for this research. Suppose there are two datasets: a target dataset (T) and a source dataset (S), and dataset T can have a different number of instances in each class. Further assume that datasets S and T have the same classes of the sentiment analysis. The goal of instance transfer involves the following process.

In order to achieve the training task of sentiment classification model in T, it chooses the transferable instances of same class from S and transfers them to the corresponding class in dataset T to create a new target dataset D', while it ensures that different classes in dataset D' have a similar data size. This helps to improve the performance of the classification model that is trained on dataset D'. The figure shows that both of datasets T and S have two same classes to be recognized, known as Pos (Positive) and Neg (Negative). After instance transfer, the instances of these classes in new dataset D' have a similar number.

The challenges of implementing this core idea are as follows: (1) how to measure the transferability of instances in S, and (2) how to homogenize the feature space of these instances with that of T. The similarity between feature space $\Omega(F|T)$ in T and feature space $\Omega(F|S)$ in S is adopted to evaluate the transferability of each instance in S. If $\Omega(F|T) = \Omega(F|S)$, then instance transfer becomes a simple task to be solved as they have direct transferability. However, in general, datasets S and T not only have common words in the unigram sets or phases in the bigram set, but also have their own innate and unique words in the unigram set or phases in the bigram set. This leads to the issue of feature space inconsistency between T and S which can be represented as $\Omega(F|T) \neq \Omega(F|S)$.

We use two datasets collected from two famous Chinese e-commerce portals, Jingdong (www.jd.com) and Dangdang (www.dangdang.com), and are named as JingDong and Dangdang, respectively in this research. The feature space of both datasets is one or many types of N-gram features, such as the unigram and

bigram of the product reviews corpora. In these two corpora, the products (as topics) of Jingdong only include Laptop and PC, while the topics of Dangdang only includes digital product accessories. The number of items of the unigram and bigram in the feature sets, JingDong and DangDang, are 1,385 and 1,258, and most of the items are different.

Inspired by the idea of topic sentences, Baxendale (1958) and Paice (1980) provide a strong indication of overall subject in each product review, this research proposes a topic sentence-based instance transfer method for imbalanced emotion classification of Chinese product reviews. The contributions of the proposed approach are as follows:

(1) Introduce a concept topic sentence for each product review. An algorithm for identifying a topic sentence for each product review is proposed based on features of title, first sentence or last sentence of the review

(2) Introduce new feature spaces, based on two feature sets, features of topic sentences and features of the whole body of each review. A *feature set* of a topic sentence includes syntax features and the frequency of emotion words and relevant nouns, as shown in Table 1.

INSERT TABLE 1 ABOUT HERE

(3) Propose a feature selection strategy for transferable instances, which is a greedy algorithm based on a function of extracting the proportion of sum of the information gain of top-N common features between the T and S datasets. This strategy helps to choose a set of common features, which contribute towards improvement of imbalanced data classification.

(4) Introduce a Smote-based method (Chawla 2003) for processing feature space inconsistency in order to overcome the inconsistency problem between feature spaces of T and the instances transferred from dataset S.

(5) Generate a training dataset by immigrating instances depending on emotion class distribution of both T and S.

Note that, the datasets we used contain two similar scales of minority emotion classes. The terms, sentiment and emotion are interchangeable, and there is no difference between them in this article (Tian et al. 2014).

2. RELATED WORKS

In the field of sentiment analysis of product reviews, two important issues, such as feature selection and classification methods, need to be discussed.

Different features for sentiment classification are used to analyze product reviews (e.g., Wu et al. 2009, Hu and Liu 2004, Sharma et al. 2014, Archak et al. 2007, Pang and Lee 2008, Kang et a. 2012, Cho et al. 2014). Chen and Tao (2010) use dependency parsing with shallow semantic analysis for Chinese opinion related expression extraction. Wu et al. (2009) use phrase dependency parsing for opinion mining. Hu et al. (2004) used frequent item sets to extract the most relevant features from a domain and pruned it to obtain a subset of features, while abstracted the nearby adjectives to a feature as an opinion word regarding that feature. Kang et al. (2012) adopted sentiment unigrams and bigrams as features, and N-grams are also used (Zhang et al. 2011). Mukherjee and Bhattacharyya (2012) abstract parts of speech tags, all nouns, direct neighbors and dependency relationships as the space of product feature. Cho et al. (2014) presented a data-driven method for adapting sentiment dictionaries to diverse domains and showed that the integrated sentiment dictionary constructed using “merge,” “remove,” and “switch” operations robustly outperforms individual dictionaries in the sentiment classification. Fu et al. (2013) adopted HowNet lexicon for sentiment analysis of product reviews.

Currently, different kinds of data mining based techniques are employed in sentiment analysis of product

reviews. Liu et al. (2013) applied text mining and natural language processing (NLP) approach to design NLP rule-based models for predicting sentiments in test data consisting of six hundred textual reviews for each app from Google Play and the Android App Store. Mukherjee and Bhattacharyya (2012) developed a system (rule-based and supervised classification) that extracts potential features from a review and clusters opinion expressions describing each of the features, which achieves a high accuracy across all domains and performs at par with state-of-the-art systems. Albornoz et al. (2011) proposed a feature-driven approach for product review rating, and their proposed joint model based method performs significantly better than the previous approaches on featuring 1,000 hotel reviews from Booking.com. Maks and Vossen (2013) incorporated standard machine learning techniques, such as Naïve Bayes and SVM, into the domain of Cantonese online restaurant reviews to automatically classify user reviews as positive or negative. Kang et al. (2012) proposed an improved Naïve Bayes algorithm for sentiment analysis of restaurant reviews and got a higher accuracy than the original Naïve Bayes and support vector machine (SVM). Three supervised machine learning algorithms, Naïve Bayes, SVM and character based N-gram model are adopted for sentiment classification in Ye et al. (2009). Recently, Wang et al. (2014) proposed a semi-supervised deep learning model that introduces supervised sentiment labels into traditional neural network language models for sentiment analysis. Both Fu et al. (2013) and Bagheri et al. (2013) adopted unsupervised methods for sentiment analysis of product reviews. After analyzing related literatures, we conclude that most of the aforementioned methods are based on supervision approaches and only balanced datasets are used in their models.

Imbalanced data classification is a challenging problem in the field of machine learning (He and Ma 2013). 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 (Barandela et al. 2004). The common methods to handle this problem include data level

sampling, cost sensitive learning, feature selection, feature weight adjustment and one-class learning (Ogura et al. 2011, Satyam and Sanjeev 2011).

Data level sampling mainly contains two basic methods known as *over-sampling* and *under-sampling*. Under-sampling extracts some data from majority class to balance the class distribution. Over-sampling repeatedly samples the minority class or directly copy them to increase the size of minority class to balance the class distribution. Pan et al. (2010) and Barandela et al. (2004) discuss advantages and disadvantages of these two sampling methods in relation to handling imbalanced problem. Under-sampling leads to data loss, while over-sampling increases training time and causes the effect of over-fitting.

The main idea of cost sensitive learning is to assign different weights to elements in a fusion matrix of classified results when the instances of minority class and majority class are misclassified, which forces the classifier to pay more attention to minority class. Kamel et al. (2007) proposed a boosting method based on cost sensitive training. Zhou et al. (2006) suggested another method that adopts a neural network for cost sensitive learning to handle the imbalance problem.

The idea behind feature selection is to choose features that are biased towards minority class in order to improve the learning outcome of minority class. Ogura et al. (2011) proposed three metrics to select features, which are biased towards minority class. They pointed out that these three metrics should be used synthetically. Liao and Pan (2012) proposed a method that selects features biased towards minority class by using feature distribution information. Wang et al. (2014) emphasized the problem of sentiment classification on imbalanced data and proposed a boundary region cutting algorithm that is only suitable for two-category sentiment classification problems, and rely on a single dataset.

The feature weight adjustment corrects the classifier bias by assigning a higher weight to features that is more important to minority class to solve the imbalance problem. Liu et al. (2009) proposed a

method that adjusts feature weights according to a distribution ratio of the minority class and the majority class to increase the influence of minority class.

One-class learning is mainly applied to situations in which the class distribution is seriously imbalanced, such as information filtering and fraud detection. One-class learning trains a model by using a single class and ignores other information. Raskutti and Kowalczyk (2004) investigated the limitation of two-class discrimination from the data with heavily unbalanced class proportions. They pointed out that there is a consistent pattern of performance differences between one-class and two-class learning for all SVMs.

These research efforts solve imbalanced problem aimed at a single target data set. These efforts make full use of the information in the data to solve the problem. In recent years, researchers have begun to adopt auxiliary datasets to solve the classification problem in different applications (Nguyen et al. 2011, Tommasi and Tuytelaars 2014, Gong et al. 2012, Heim et al. 2014, Hung and Lin 2011, Pan and Yang 2010). The present work is inspired by the idea of topic sentence and aims to transfer similar instances from auxiliary datasets into a target dataset in order to overcome the imbalanced class distribution problem.

3. A TOPIC SENTENCE-BASED INSTANCE TRANSFER METHOD

As we have noted, the challenge for the instance transfer method is how to measure the transferability of the instances (e.g., product reviews) in a dataset S. A top priority task of measuring the transferability of the instances is to find common features between T and S. As we understand them, online product reviews tend to have a kind of paragraph-like writing-style. Inspired by the concept of topic sentences used in the automatic generation of abstracts of literatures (Baxendale 1958), we will apply the similarity of the topic sentences of two product reviews in different data sets to measure their similarity. A topic sentence essentially tells what the rest of the paragraph is about. Note that the meaning of *topic* in *topic sentences* is different from the meaning of topic *modelling*. The topic in the field of topic modelling (Tian et al. 2014) is an

object (e.g., a product), event or domain, while a topic sentence gives a strong indication of its overall subject (Paice 1980).

Moreover, the core idea behind a common-feature selection-based instance transfer method is as follows: considering that the classification task on datasets S and T is same, we denote the feature space in T and the one in S as $\Omega(F|T)$ and $\Omega(F|S)$ respectively, and then transfer similar instances in S into T. In general, $\Omega(F|T) \neq \Omega(F|S)$. In this article, including the feature set of topic sentences, the features of product reviews have syntactic features, frequency features and N-gram features. (See Table 1 again). In syntactic features, a Chinese sentiment lexicon base is adopted, which includes HowNet and others that were manually collected in our prior works (Tian et al. 2011, 2014). The N-gram feature refers to the combinations of the words and has a strong dependency on data/corpus. In this article, Bigram and Unigram are two feature subsets of N-gram.

Based on the topic sentence, the challenges to implement the core idea are how to identify a topic sentence of each product review and evaluate the similarity and effectiveness of $\Omega(F|T)$ in T and $\Omega(F|S)$ in S, and how to overcome the inconsistent feature space between T and S that is caused by their unique features. We should solve the following problems: (1) identifying a topic-sentence of each product review and abstracting its features; (2) discovering and selecting common features of T and S; (3) evaluating the transferability of each instance in dataset S; and (4) homogenizing incoherent feature spaces between transferred instances and dataset T to overcome issue of feature space inconsistency.

This article proposes a new approach to solve the above problems. The frame diagram of the approach is shown in Figure 2.

INSERT FIGURE 2 ABOUT HERE

The dataset T contains N1 pieces of review instances and dataset S contains N2 pieces of review in-

stances. F_T represents a matrix of the feature values of dataset T and has k dimensions common features and p dimensions N-gram features. F_{STP} represents a matrix of the feature values of the dataset that contains M2 pieces of transferable instances, and has k dimensions common features and p dimensions 0-value. F_{Homo} represents matrix of the feature values of the dataset that contains M2 pieces of transferable instances, and has k dimensions common features and p dimensions N-gram features generated. Matrix F_{NEW} is the union of F_T and F_{Homo}. The approach encompasses five steps:

- **Step 1: Topic-sentence identification**. This step corresponds to the label ① in Figure 2. A topic sentence of each product review is identified according to position and content of the sentences in each product review.
- **Step 2: Common feature selection**. This step corresponds to the label ②. A greedy algorithm based on a function for calculating proportion of sum of the information gain of Top-N common features of topic sentences between T and S is employed to solve the problem of discovering and selecting common features. In this article, common features are used to represent common features of topic sentences.
- **Step 3: Transferability evaluation**. This step corresponds to the label ③. It evaluates the transferability of each instance in dataset S to determine appropriate instances to transfer. It can be divided into two sub-problems: (1) Determining a suitable amount of the transferred instances; (2) choosing appropriate instances from dataset S. To solve sub-problem 1, it starts with balancing the instance size of the minority class in T to overcome its class imbalance. For the sub-problem 2, we adopt the cosine similarity scores based on common features of topic sentences to measure the similarity between instances in S and the corresponding ones in T.
- **Step 4: Homogenization**. This step corresponds to the label ④. It involves processing of the fea-

ture 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: New dataset and Training**. This step corresponds to the label ⑤. It immigrates the transferable instances in S into dataset T by considering different emotions in order to form a new target dataset *D'*, and it trains different classifiers on it and evaluate and compare their performances on the trained classification models to select the best one.

The following subsections describe the proposed method in detail. We first describe a topic-sentence identification method. Then we explain the method of selecting the common features of both T and S. Next we present a cosine similarity calculation method for selecting the transferable instances from source dataset, which measures the transferability of each instance in S. And last, we introduce the homogenization process for the feature space of transferable instances in S.

3.1. Topic-sentence identification method

We investigated the collected data and discovered that, if there is a title of a product review or it has only one sentence in a product review, it is definitely a topic sentence. Otherwise, most of the time, the topic sentence of the product review is located in the first or last sentence. So, a rule and supervision learning hybrid method for identifying topic sentence is proposed. To wit, if there is a title of a product review or there is only one sentence in a product review, the method labels it as a topic sentence of the product review. Otherwise, we use nouns (e.g., product name, type and its producer), their frequency of occurrence in the review, relevant keywords of products, emotion words and their POS-tags, and their dependency in the first sentence and last sentence of each product review to form the feature set.

We apply seven classification algorithms: J48, Random Forest, ADTree, AdaBoostM1, Bagging, Mul-

tilayer Perceptron and Naïve Bayes. These are used to label the datasets while applying ten-fold cross-validation to test the performance of each classification model. For the experiment, we used the Bagging method to assess whether a topic sentence was either the first or last sentence in a product review.

3.2. Common features selection in the source and target datasets

In the feature set for topic sentences, category variables are majority variables. After computing, we found that the proportion of sum of the information gain of common features between T and S has a relatively large proportion in both datasets (Han and Kamber 2006). So, we decided to utilize this proportion to select common features. The steps of this process are as follows:

(1) 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.

(2) Mark the position of common features in the sorted list.

(3) For each marked position, compute the proportion of the sum of information gain of common features at the specific position and all other features lower than that position and the sum of information gain of all the features which appear before the position. This is the *proportion of the sum of the information gain of common features between* T *and* S. Select the common features that have larger proportions to construct the feature set to represent the instances.

This process is shown in Figure 3, which is used for evaluating features by considering their weights in both datasets.

INSERT FIGURE 3 ABOUT HERE

The element position in the two different ranked lists shows the difference of their importance in classification. There is a subset of common features of T and S before the position of each element in the common features. The sum of the weight of this subset before the element's position reflects the importance

of this subset. In Figure 3, F_S represents the feature set of dataset S, and R_1 is the dimension of F_S; F_T represents the feature set of dataset T, and R_2 is the dimension of F_T; $R_1 \neq R_2$; $F_{S\cap T}^{com}$ represents the common features of S and T datasets. Also, R is the dimension of $F_{S\cap T}^{com}$.

In the computational process, the function *mode* calculates the element number in each dataset. If $total \neq \varnothing$ and the S and T datasets have no common feature, the algorithm stops. $F_S^{'}$ is the feature set of S dataset, and the features in it have been arranged in descending order based on the information gain. $F_T^{'}$ is the feature set of dataset T. Its features have been arranged in descending order of their information gain. Function $IG(F)$ calculates the information gain of each feature in the feature space of corresponding dataset. Function *Sort* is to rank the data in descending order according to specified value. The equation $[F_{S\cap T}^{'}, index_{F_S}] = \bigcap(F_S^{'}, F_T^{'})$ is used to find the common features of S and T and return numerical value $index_{F_S}$. $F_{S\cap T}^{'} = \{f_{S\cap T}^{'1}, f_{S\cap T}^{'2}, f_{S\cap T}^{'3}, ... f_{S\cap T}^{'M}\} = \{f_{S\cap T}^{'1} \mid i = 1,2,...M\}$, $M = \mathrm{mod}\, e(F_{S\cap T}^{'})$ and $M == total$; $index_{F_S} = \{index(g) \mid g = 1,2,...M\}$ represents the index of the common features of S and T datasets in $F_S^{'}$; $\max\_index = Max(TopN)$ is used to find the features which have the largest proportion of sum of the information gain of Top-N common features between target and source datasets, and return its index in $F_S^{'}$. In line 21, the feature set $F_{S\cap T}^{com}$ is obtained.

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

Cosine similarity is a common method for calculating two files' similarity in natural language processing, in which each file is represented as a feature vector. This research adopts 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 three steps:

- **Step 1:** Express each instance with the selected common features in a vector form, and normalize them. The feature normalization process involves two sub-steps: 1, processing category attributes:

All category attributes/features are replaced directly with numerical value starting from 0 and increased by 1 subsequently. For example, the feature conjunction has 8 values: none, turn, casual, supposition, coordinate, comparison, undertake and select. 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 to normalize numerical features (Tian et al. 2014).

- **Step 2:** Calculate the overall cosine similarity scores between corresponding emotion instances from source dataset and the emotion instances in target dataset. Generally, the more similar two instances are, the higher their overall cosine similarity score is. $L = \{l_1, l_2, \ldots, l_N\} = \{l_p \mid p = 1, 2, \ldots, N\}$ denotes a set of class labels, N denotes the number of labels of classification tasks. Here, $N = 2$, l_1 represents positive emotion, and l_2 represents negative emotion. The formula of cosine similarity calculation is:

$$score(InsSou^{l_p}(i)) = \frac{\sum_{i=1}^{m} COS(InsSou^{l_p}(i), InsTar^{l_p}(j))}{m} \quad (1)$$

where $InsTar^{l_p}(j))$ denotes an instance labeled with l_p in the target dataset. Here, $j = 1,2, \ldots, n$ indicates that there are n instances with the same label in the target dataset. $InsSou^{l_p}(i))$ denotes an instance labeled with l_p in the source dataset, and $i = 1,2, \ldots, K$ denotes that there are K instances with the same label in the source dataset. $COS(InsSou^{l_p}(i), InsTar^{l_p}(j))$ means the common features-based cosine similarity score between $InsTar^{l_p}(j))$ and $InsSou^{l_p}(i))$, 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 the same labels 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.

3.4. Homogenization processing of the 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 a 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 numerical. We adopted the Smote method to produce the values of N-gram features of each instance to be transferred in order to make transferable instances compatible with the target dataset.

3.5. Instance combination and model training

We have provided details of 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 one to overcome the imbalanced problem in the target dataset. The next step is to train a sentiment classification model. The instance combination conforms to following two principles:

(1) An instance can only be transferred once, the reason is that multiple transfer of one same instance will cause over-fitting problem.

(2) Make the number of instances in each emotion class in T balanced, to overcome the imbalance in the target dataset as much as possible.

4. EXPERIMENTS AND THEIR ANALYSIS

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

4.1. Experiment

The experiments involve these steps:

- **Step 1: Collect experimental corpora.** Two datasets were collected from two famous Chinese e-commerce portal, Jingdong and Dangdang. The feature space of both datasets are one or many types of N-gram features, such as Unigram and Bigram, of the product review corpora, as well as the manually collected sentiment word base (Tian et al. 2012) is adopted when abstracting the features. In both corpora, the topics (products) of Jingdong only include Laptop and PC, while the topics of Dangdang only includes digital product accessories. Each review and its topic sentence in these corpora are labeled manually with polarity, negative or positive. Features (as shown in Table 1) and N-gram (bigram and unigram according to TF-IDF (term frequency-inverse document frequency)) are abstracted from Jingdong and produce two datasets, JDTSF and JDN-gram. Combining JDTSF and JDN-gram forms a new dataset JD. After abstracting these two features from Dangdang, we obtain DDTSF and DDN-gram. Merging the two datasets forms a new dataset DD. JDTSF, JDN-gram and JD are imbalanced datasets, while DDTSF, DDN-gram and DD are balanced datasets. So we take JD as the target dataset and DD as the source dataset.
- **Step 2:** Identify the topic sentence of each review in Jingdong by employing the method described in Section 3.1. on how to evaluate the performance of the proposed methods.
- **Step 3: Select common features of topic sentences.** Based on JDTSF and DDTSF, we select the common features of topic sentences according to the steps mentioned in Section 3.2 and calculate the overall cosine similarity between instances in source dataset and instances in target dataset. This

enables us to determine the instances to be transferred from the source dataset.

- **Step 4: Carry out feature space homogenization.** This involves applying 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 target dataset.** This is done according to the steps described in Section 3.4, in order to form a new training dataset, JDImmigration. Note that for comparison with traditional data sampling strategies and methods for imbalanced datasets, two other datasets, JDResample and JDSmote, also area produced by applying resampling and Smote to JD.
- **Step 6: Apply the five classification algorithms.** These include: J48, Random Forest, SVM, Random Committee and Naive Bayes to the above datasets, while using ten-fold cross-validation to test the performance of each classification model. Note that "RF" denotes the Random Forest classification algorithm, "SVM" is the support vector machine method (Platt 2008) "RC" denotes the Random Committee classification algorithm, and "NB" denotes the Naive Bayes classification algorithm. For the classification models, we use JDN-Gam, JD, JDResample, JDSmote, and CFImmigration as the training dataset for the classification method and an extra training dataset, JD634, for which we collected 634 instances from Jingdong.

In the classification experiments, "P", "R" and "F" denote precision, recall and the F1-measure respectively. *Precision i*s the ratio of the classified relevant instances divided by all classified instances, while *recall* (also known as *sensitivity*) is the ratio of all classified relevant instances divided by all relevant instances in the dataset. The *F1-measure* is the harmonic mean of precision and recall. The classification experiments were carried out using Weka (Hall et al. 2009). In addition, the weighted average of each in-

dicator in our experiment is the result of multiplying the value of the indicator in each emotion class (Pos, Neg) by corresponding weights and adding the sum of the overall value, then dividing the total sum by total number of units.

4.2. Experimental results

After carrying out Step 1, the number of features in the feature sets of JD and DD were 1418 and 1291, respectively. The numbers of N-grams in the two datasets were 1,385 and 1,258, respectively. The number of Pos instances and Neg instances in DD was 2,887. The number of Pos instances in JD was 1,600 while the number of negative instances in JD was 320.

INSERT TABLE 2 ABOUT HERE

Table 2 shows the weighted averages of the precision, recall and F1-measures for seven classification algorithms on the identification of topic sentence. After executing the method described in Section 3.1, the average accuracy of identifying the topic sentence of each review of JD was 87.8%. The Bagging algorithm showed the best performance. The common features were selected by applying the method described in Section 3.2.

After executing Step 4, the number of the transferred instances from DD was 1,280 to make JDImmigration balanced. Thus, the number of both positive and negative instances in JDImmigration was 1,600.

To highlight the overall performance, we listed and analyzed the weighted averages of the precision, recall and F1-measure. We explain performance related to Pos and Neg emotions also, but related experimental results are shown in tables in an Appendix at the end.

Figures 4 to 6 show part of experimental results corresponding to Steps 6 and 7 in our experiments.

INSERT FIGURES 4, 5 and 6 ABOUT HERE

Figure 4 shows the weighted average of precision for ten-fold cross-validation and the generalization ability evaluation. The dotted lines depict the cross-validation results when applying J48, NB, RC, RF and SVM on JDN-gram, JDTSF, JD, JDResample, JDSmote and JDImmigration. The solid lines show the results for generalization ability when applying J48, NB, RC, RF and SVM on JDN-gram, JD, JDResample, JDSmote and JDImmigration.

The five dotted lines show that the best results for the weighted average of precision with ten-fold cross-validation were achieved by applying RF to JDResample, which achieves a value of 0.99. The four methods, NB, SVM, RF and RC applied on JD, JDResample and JDSmote perform better than when they applied on JDN-gram. Compared with the performance achieved on JDN-gram, the average improvement sin weighted average of precision of JD, JDResample, JDSmote and JDImmigration were 0.99%, 4.40%, 1.94% and 2.14%, respectively.

According to the five solid lines in Figure 4, the best result for the weighted average of precision in generalization ability was achieved by applying SVM to JDImmigration, and the related value is 0.913. Compared with the performance achieved on JDN-gram, the average performance improvement in the weighted averages of the precision of JD, JDResample, JDSmote and JDImmigration were 1.59%, 2.63%, 0.69% and 5.50%, respectively. This shows that our proposed method helps the adopted classification algorithms perform better than other methods for this assessment.

Note that the percentage of average improvement is equal to the average of the difference of the five methods' performance on JDN-gram and the other datasets. The percentages of average improvements mentioned in the following paragraphs were calculated in the same way.

Figure 5 shows the weighted averages of recall for ten-fold cross-validation and the assessment of generalization ability. The dotted lines depict the results of ten-fold cross-validation when applying J48, NB,

RC, RF and SVM on JDN-gram, JDTSF, JD, JDResample, JDSmote and JDImmigration. The solid lines show the results of generalization ability when conducting J48, NB, RC, RF and SVM on JDN-gram, JD, JDResample, JDSmote and JDImmigration.

According to five dotted lines, the best result of weighted average of recall in ten-fold cross-validation is achieved by applying RF to JDResample, for a value of 0.99. The four methods, NB, SVM, RF and RC applied on JDResample and JDSmote, performed better than JDN-gram. And, compared with the performance achieved for JDN-gram, the average improvements in the weighted average of recall of JD, JDResample, JDSmote and JDImmigration were 1.44%, 5.05%, 2.50% and 1.96%, respectively.

As shown by the solid lines, the best result of weighted average of recall in generalization ability evaluation was achieved by applying SVM to JDImmigration, and a value of 0.907 was achieved. The four methods, J48, RC, RF and SVM applied on JD, JDResample, JDImmigration and JDSmote performed better than for JDN-gram. Compared with the performance on JDN-gram, the average performance improvements in weighted average of recall of JD, JDResample, JDSmote and JDImmigration were 11.11%, 15.91%, 10.32% and 23.91%, respectively. This shows that our proposed method helps the adopted classification algorithms perform much better than the others for the conditions that we tested.

We next consider the weighted average of the F1-measure of ten-fold cross-validation and the generalization ability evaluation. The dotted lines describe the results of ten-fold cross-validation when applying J48, NB, RC, RF and SVM on JDN-gram, JDTSF, JD, JDResample, JDSmote and JDImmigration. Similarly, the solid lines show the results of generalization ability when applying J48, NB, RC, RF and SVM on JDN-gram, JD, JDResample, JDSmote and JDImmigration.

The five dotted lines shown in Figure 6 indicate that best result for the weighted average of the F1-measure in ten-fold cross-validation was achieved by applying RF to JDResample, and returned a value

of 0.99. The five methods, J48, NB, SVM, RF and RC applied on JDResample, performed better than when they were applied on JDN-gram. Compared with the performance on JDN-gram, the average improvements in the weighted average of F1-measure of JD, JDResample, JDSmote and JDImmigration were 1.54%, 5.82%, 3.08% and 2.58%, respectively.

According to five solid lines, the best result for the weighted average of the F1-measure related to the generalization ability evaluation was achieved by applying SVM to JDImmigration. The resulting value of was 0.908. The five methods applied on JD and JDResample performed better than when they were applied on JDN-gram. Compared with the performance achieved on JDN-gram, the average improvements in the weighted averages of the F1-measure of JD, JDResample, JDSmote and JDImmigration were 14.18%, 20.0%, 14.14% and 28.54%. This shows that our method helped the adopted classification algorithm perform better than Resample and Smote under the test conditions.

Based on the experimental results, their analysis and the additional tables that are shown in the Appendix, our conclusions are:

(1) As can be observed from dotted lines shown in Figure 4-6 and Tables A1 to A9 that the performance of four classification methods (J48, RC, RF and SMV) when applied on the feature set of topic sentences, JDTSF seems to be good enough to achieve performance comparable to JDN-gram. We can draw this conclusion based on the weighted averages of the precision, recall and F1-measures. This verifies that a topic sentence is a strong indication of the overall subject in each product review.

(2) In Step 6 of the experiment, for ten-fold cross-validation, JDResample had the best performance. It achieved 0.99 for all of the precision, recall and F1-measures. The high performance of JDResample was mainly due to over-fitting. This was caused by applying the Resample method to the minority class of JD. The number of its negative instances increased from 189 to 1,880. Repeatedly sampling

and the classification models trained on this kind dataset tend to memorize information on the features of of the duplicated instances in the minority class of JD. The same is true for JDSmote.

(3) In order to evaluate the ability of generalization of Resample, Smote and our proposed method, we conducted Step 7 of our experiment. The results show that our method has a stable improvement in the recall and F1-measure when applying J48, SVM, RC and RF. Moreover, the F1-measure is the most commonly used method to comprehensively consider the precision and recall indicators. It effectively reflects the performance of the classification methods. Therefore, comparing the weighted average sof F1-measure in the experiment, the average improvements from applying the four classification methods (SVM, RF, RC and J48) to the immigration dataset produced by our method outperformed the other methods. Its results for the generalization ability evaluation as well as most of the results for the ten-fold cross-validation reflected performance improvements. This suggests that our method overcomes the influence of over-fitting the data and does well in terms of its ability for generalization in terms of the weighted averages of the F1-measure.

(4) Considering the weighted performance indices that we used in the experiments for evaluating the ability of generalization, using SVM on JDImmigration outperformed the other classification methods when applied to datasets produced by Resample and Smote.

(5) Finally, in our experiments, the improvement of performance for negative emotions came only with some sacrifice of performance related to positive emotions. This can be observed from the results that are dispayed in Tables A1 to A18. According to Tables A4 to A6 and Tables A12 to A15 especially, the immigration dataset produced by our proposed method improved the classification performance for the minority class of negative emotion. This was significant for both the ten-fold cross-validation and generalization ability evaluation.

5. CONCLUSION

To effectively address the challenge of imbalanced sentiment analysis of product reviews, this article proposes a topic sentence-based instance transfer method. This method is inspired by the topic sentence and combines a feature set of topic sentence with N-gram features as the new feature set. Firstly, a rule and supervised learning hybrid method is designed to identify topic sentence of a product review. Secondly, after incorporating the feature set of the topic sentence into the feature space of sentiment classification, a greedy algorithm based on a function of extracting the proportion of sum of the information gain of top-N common features between source dataset and target dataset is proposed to help select the transferable instances. Next, a SMOTE-based method for processing feature space inconsistency in order to overcome the inconsistency problem between feature spaces of T and the instances transferred from dataset S. Extensive experiments on different datasets produced by N-gram, resample, Smote and our proposed method are carried out. The experimental results show that (1) with the help of newly added features of topic sentence, many methods perform better than as on N-gram features; (2) it can be verified that resample leads to over-fitting problem of the trained classification model; (3) the most importantly in the experiments for evaluating the ability of generalization, SVM outperforms J48, Random forest, Random Committee and Naive Bayes according to the weighted average of performance indices, precision, recall and F1-measure.

Future work will focus on adapting our instance transfer method to process large scale corpora, even unlabeled ones. Moreover, the long-term vision for our research is to implement and employ a reliable service for a real e-commerce platform (Immonen and Pakkala 2014, Huergo et al. 2014). The service will analyze imbalanced sentiments in product reviews in real time.

REFERENCES

Archak N, Ghose A, Ipeirotis P G (2007) Show me the money! deriving the pricing power of product features by mining consumer reviews. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, New York, NY, 56-65.

Bagheri A, Saraee M, De Jong F (2013) Care more about customers: unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. Knowledge-Based Systems, 52, 201–213.

Barandela R, Valdovinos R M, Sánchez J S (2004). The imbalanced training sample problem: under or over sampling? In Structural, Syntactic, and Statistical Pattern Recognition, Berlin, Heidelberg: Springer, Berlin Heidelberg, Germany, 806-814.

Baxendale P B (1958) Machine-made index for technical literature: an experiment. IBM Journal of Research and Development, 2, 354-361.

Chawla N V (2003) C4.5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure. Proceedings of the 2003 International Conference on Machine Learning, AAAI Press, Palo Alto, CA, 2003.

Chen M, Yao T (2010) Combining dependency parsing with shallow semantic analysis for Chinese opinion-element relation identification. 4th IEEE International Universal Communication Symposium, IEEE Computer Society Press, 299-305.

Cho H, Kim S, Lee J, Lee JS (2014) Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews. Knowledge-Based Systems, 71, 61–71.

de Albornoz J C, Plaza L, Gervás P, Diaz, A (2011) A joint model of feature mining and sentiment analysis for product review rating. Advances in Information Retrieval, Lecture Notes in Computer Science, 6611, Springer, Berlin Heidelberg, Germany, 55-66.

Fu X, Liu G, Guo Y, Guo WB (2013) Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. Knowledge-Based Systems, 37, 186–195.

Gong B, Sha F, Grauman K. (2012) Overcoming dataset bias: an unsupervised domain adaptation approach. Big Data Meets Computer Vision: First International Workshop on Large Scale Visual Recognition and Retrieval, Lake Tahoe, NV.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update, SIGKDD Explorations, 11(1), 10-18.

Han J, Kamber M. (2006) Data Mining: Concept and Techniques, 2nd Ed. Morgan Kaufmann , San Francisco, CA.

He H, Ma Y (2013) Imbalanced Learning: Foundations, Algorithms and Applications. IEEE Computer Society Press, Washington, DC.

Heim E, Valizadegan H, Hauskrecht M (2014) Relative comparison kernel learning with auxiliary kernels. Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science, 8724, 563-578.

Hu M, Liu B. (2004) Mining and summarizing customer reviews. In proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Dta Mining. ACM Press, New York, NY, 168-177.

Hung C W, Lin H T (2011) Multi-label active learning with auxiliary learner. In Proceedings of the Asian Conference on Machine Learning, in JMLR Workshop and Conference Proceedings, 20, 315-330.

Huergo R S, Pires P F, Delicato F C, Costa B, Cavalcante E, Batista, T (2014) A systematic survey of service identification methods. Service Oriented Computing and Applications, 28 (3), 199–219.

Immonen A, Pakkala D (2014) A survey of methods and approaches for reliable dynamic service compositions. Service Oriented Computing and Applications, 8(2), 129-158.

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

Kang H, Yoo S J, Han D (2012) Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Systems with Applications, 39, 6000–6010.

Liao Y, Pan X (2012) Feature selection method on imbalanced text. Journal of Xidian University, 41(4), 592-595.

Liu Y, Loh H T, Sun A (2009) Imbalanced text classification: A term weighting approach. Expert systems with Applications, 36(1), 690-701.

Liu J, Sarkar M K, Chakraborty G. (2013) Feature-based sentiment analysis on Android app reviews using SAS® Text Miner and SAS® Sentiment Analysis Studio. In Proceedings of the SAS Global Forum 2013 Conference, San Francisco, CA, May 1, 250.

Maks I, Vossen P (2013) Sentiment analysis of reviews: should we analyze writer intentions or reader perceptions? Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria, 415–419.

Mukherjee S, Bhattacharyya P (2012) Feature specific sentiment analysis for product reviews. Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science, 7181, Springer Berlin Heidelberg, 475-487.

Nguyen Q, Valizadegan H, Hauskrecht M (2011) Learning classification with auxiliary probabilistic information. In Proceedings of the 2011 IEEE International Conference on Data Mining, IEEE Computing Society Press, Washington, DC, 477–486.

Ogura H, Amano H, Kondo M (2011) Comparison of metrics for feature selection in imbalanced text classification. Expert Systems with Applications, 38(5), 4978-4989.

Paice C D (1980) The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval, Butterworth, Kent, UK, 172-191.

Pan S J, Yang Q (2010) A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345-1359.

Pang B, Lee L (2008) Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 459-526.

Platt J (1998) Fast training of support vector machines using sequential minimal optimization. In Scholkopf B, Berges, CJC, Smola, AJ (eds.), Advances in Kernel Methods: Support Vector Learning, MIT Press, Cambridge, MA, 185-208.

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

Satyam M, J A, Sanjeev S (2011) Comparison of metrics for feature selection in imbalanced text classification. Expert Systems with Applications, 38(5), 4978-4989. Sharma R, Nigam S, Jain R (2014) Mining of product reviews at aspect level. International Journal in Foundations of Computer Science and Technology, 4(3), 87-95.

Tian F, Gao P, Li L, Zhang W, Liang H, Qian Y, Zhao R (2014) Recognizing and regulating e-learners' emotions based on interactive Chinese texts in e-learning systems. Knowledge-Based Systems, 55, 148-164.

Tian F, Liang H, Li L, Zheng Q (2012) Sentiment classification in turn-level interactive chinese texts of e-learning applications. In Proceedings of the 12th IEEE International Conference on Advanced Learning Technologies, IEEE Computer Society Press, Washington, DC, 480-484.

Tommasi T, Tuytelaars T (2014) A testbed for cross-dataset analysis. In Proceedings of the Computer Vision – ECVV Workshops, III, Zurich, Switzerland.

Ye Q, Zhang Z, Law R (2009) Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Systems with Applications, 36, 6527–6535.

Wang Y, Li Z, Liu J, Huang Y, Li D (2014) Word vector modeling for sentiment analysis of product reviews. Natural Language Processing and Chinese Computing. Springer Berlin, Heidelberg, 168-180.

Wu Y, Zhang Q, Huang X, Wu L (2009) Phrase dependency parsing for opinion mining. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 3, Association for Computational Linguistics, 1533-1541.

Zhang W, Xu H, Wan W (2012) Weakness finder: find product weakness from Chinese reviews by using aspects based sentiment analysis. Expert Systems with Applications, 39, 10283–10291.

Zhang Z, Ye Q, Zhang Z, Li YJ (2011) Sentiment classification of Internet restaurant reviews written in Cantonese. Expert Systems with Applications, 38(6), 7674-7682.

Zhang L, Hua K, Wang H, Qian, G, Zhang, L (2014) Sentiment analysis on reviews of mobile users. The 11th International Conference on Mobile Systems and Pervasive Computing, Procedia Computer Science, 34, 458-465.

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

Table 1. Feature set of each topic sentence

No.	Items of feature	Description of items of features in a topic sentence
1	negatorBlongAtt	There exists negators in the attributive part of a topic sentence
2	existDegreeBelongAtt	There exists adverbs of degree in the attributive part of a topic sentence
3	advBelongAtt	There exists adverbs in the attributive part of a topic sentence
4	adjBelongAtt	There exists adjectives in the attributive part of a topic sentence
5	existPronoun	There exists pronoun in the subjective part of a topic sentence
6	negatorBelongadverCount	Number of negators in the adverbial part of a topic sentence
7	degreeBelongAdverCount	Number of adverbs of degree in the adverbial part of a topic sentence
8	advBelongAdver	There exists adverbs in the adverbial part of a topic sentence
9	adjBelongAdver	There exists adjective in the adverbial part of a topic sentence
10	emotionVerb	There exists emotion verb in the predicate part of a topic sentence
11	nagatorBelongComplement	There exists negators in the complement part of a topic sentence
12	degreeBelongComplement	There exists adverbs of degree in the complement part of a topic sentence
13	advBelongComplement	There exists adverbs in the complement part of a topic sentence
14	adjBelongcomplement	There exists adjective in the complement part of a topic sentence
15	existObject	There exists objects in the object part of a topic sentence
16	emotionNoun	There exists objects in the object part of a topic sentence
17	sentencestructure	What topic sentence structure is, simple or clauses
18	conjunction	Conjunctions, such as casual.
19	maxEverySetence	The frequency of the most occurred character in a topic sentence
20	posWord	The frequency of positive words occurred in a topic sentence
21	negWord	The frequency of negative words occurred in a topic sentence
22	FrePunct	Frequency that a punctuation occurred in a topic sentence
23	oneWord	Frequency that a single word occurred in a topic sentence
24	twoWord	Frequency that a bigram/phrase occurred in a topic sentence
25	FreFunctionWord	The number of functional words in a topic sentence is composed of
26	FreCha	The number of characters in a topic sentence
27	FreVerb	Frequency that a verb occurred in a topic sentence
28	FreNoun	The number of nouns in a topic sentence
29	FreAdv	The number of verbs in a topic sentence
30	FreAdj	The number of adjectives in a topic sentence
31	emotionSign	Emoticons, for example, =. =, :@
32	emotionGraph	Emotional image the speaker posted.
33	otherSign	Special punctuation, for example, ??, !!, and . . . ., etc.

Table 2. Performance of applying seven classification algorithms to identify topic sentences

Classifiers	Weighted. Average		
	P	R	F
J48	**0.890**	**0.881**	**0.880**
Random Forest	0.856	0.855	0.855
ADTree	0.880	0.871	0.870
AdaBoostM1	**0.890**	0.874	0.872
Bagging	**0.890**	**0.881**	**0.880**
Multilayer Perceptron	0.853	0.853	0.852
Bayes	0.886	0.877	0.876

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

No.	Value of Information gain	Feature name
1	0.102806	adjBelongcomplement
2	0.080285	negFre
3	0.080285	posFre
4	0.079128	adjBelongAtt
5	0.077421	Function
6	0.073725	FrecharFre
7	0.058382	adjBelongAdver
8	0.057395	oneFre
9	0.05737	nounFre
10	0.052104	maxFre
11	0.043774	negatorBlongAtt
12	0.03817	otherSign
13	0.03453	nagatorBelongComplement
14	0.033746	adjFre
15	0.029066	negatorBelongadverCount
16	0.023059	twofer
17	0.018511	degreeBelongComplement
18	0.016339	emotionVerb
19	0.014923	advBelongAtt
20	0.009263	degreeBelongAdverCount
21	0.005853	emotionNoun

Figure 1. An instance transfer for imbalanced emotion classification

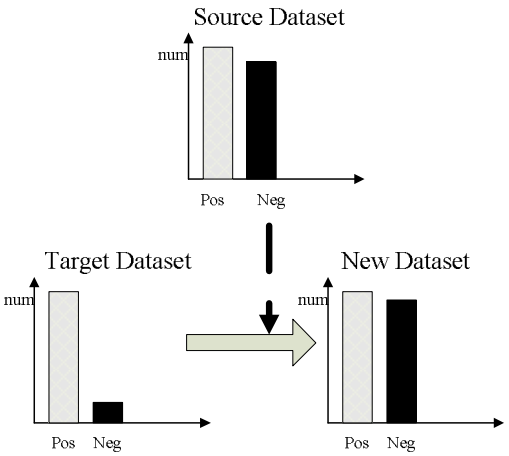


Figure 2. The frame diagram of our proposed approach

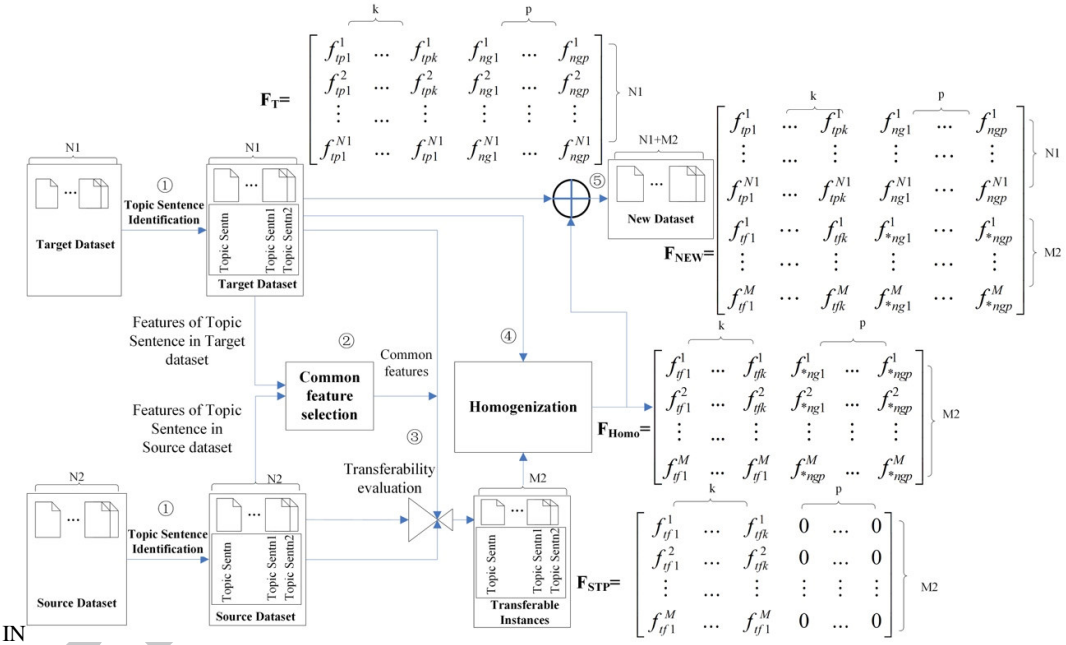


Figure 3. Pseudo code of the function of the proportion of sum of the information gain of Top-N common features between T and S

Input:

$F_s = \{f_s^1, f_s^2, \cdots, f_s^{R1}\} = \{f_s^l \mid l = 1,2,\cdots,R1\}$

$F_T = \{f_T^1, f_T^2, \cdots, f_T^{R1}\} = \{f_T^l \mid l = 1,2,\cdots,R2\}$

$F_{S\cap T}^{com} = \{f_{S\cap T}^1, f_{S\cap T}^2, \cdots f_{S\cap T}^R\} = \{f_{S\cap T}^l \mid l = 1,2,\cdots R\}$

Output:

$\max\_index = Max(TopN) < R$

$F_{S\cap T}^{'com} = \{f_{S\cap T}^{'1}, f_{S\cap T}^{'2}, \cdots f_{S\cap T}^{'\max\_index}\} = \{f_{S\cap T}^{'l} \mid l = 1,2,\cdots \max\_index\}$

Begin:

$F_{S\cap T} = F_S \bigcap F_T$

$total = \mathrm{mod}\, e(F_{S\cap T})$

IF $total = \varnothing$

break;

END

$F_S^{'} = Sort(IG(F_S))$

$F_T^{'} = Sort(IG(F_T))$

$[F_{S\cap T}^{'}, index_{F_S}] = \bigcap(F_S^{'}, F_T^{'})$

IF $i = 1:M$

$$TopN(i) = \frac{\sum_{g=1}^{index(i)} IG(f_S^{'index(g)})}{\sum_{m=1} IG(f_S^{'m})}$$

END

$\max\_index = Max(TopN)$

$F_{S\cap T}^{'com} = \{f_{S\cap T}^{'1}, f_{S\cap T}^{'2}, \cdots f_{S\cap T}^{'\max\_index}\} = \{f_{S\cap T}^{'l} \mid l = 1,2,\cdots \max\_index\}$

END

Figure 4. Weighted averages of precision for ten-fold cross-validation and generalization ability evaluation

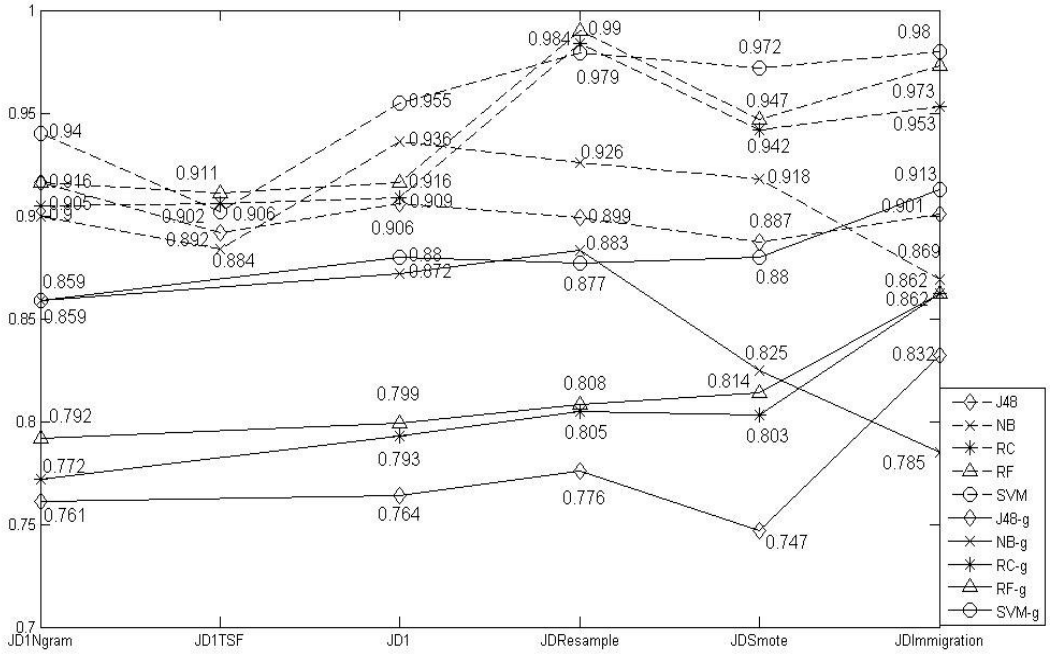


Figure 5. Weighted averages of recall for ten-fold cross-validation and generalization ability evaluation

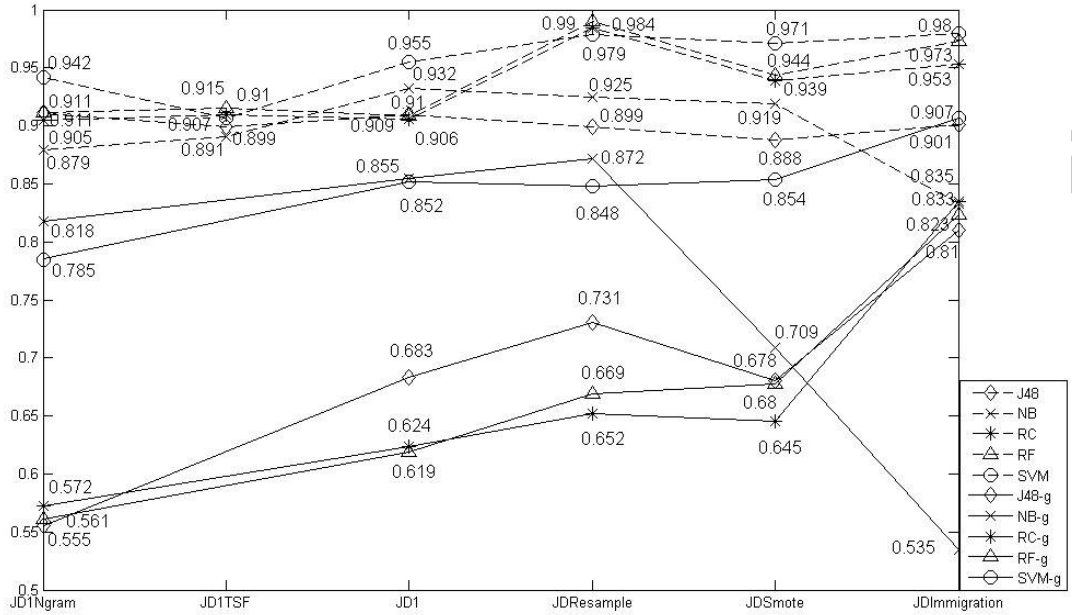
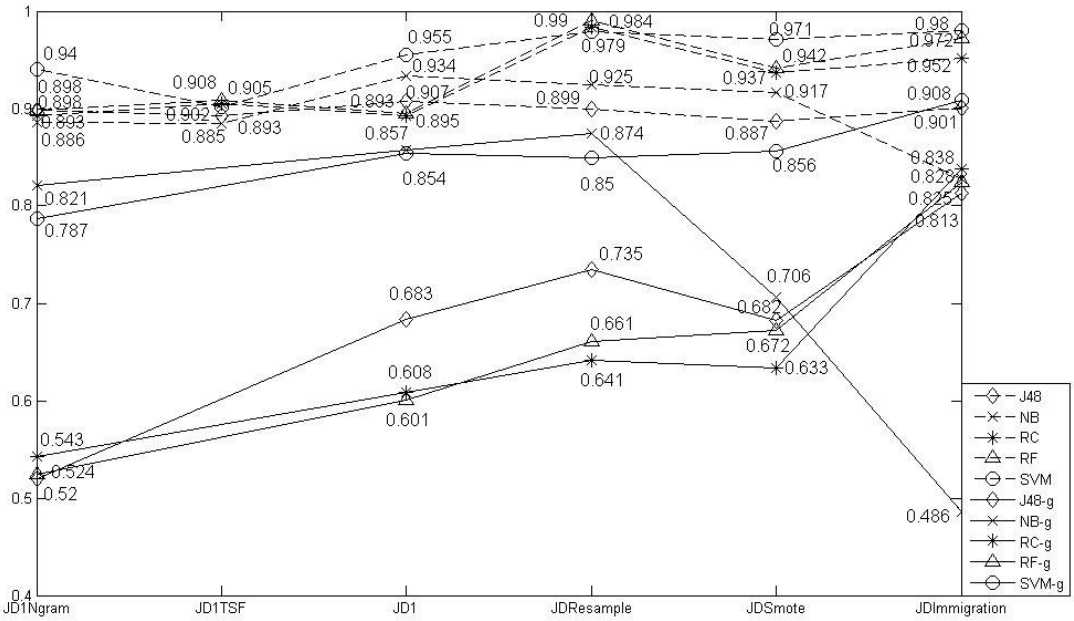


Figure 6. Weighted averages of the F1-measure for ten-fold cross-validation and generalization ability evaluation



APPENDIX. ADDITIONAL EXPERIMENTAL RESULTS

This appendix describes some experimental results that are not shown in the main body of this article.

Table A1. Precision of five methods' recognizing positive emotion on JDN-gram, JDTSF, JD, JDResample, JDSmote and JDImmigration

Positive	JDN-gram	JDTSF	JD	JDResample	JDSmote	JDImmigration
J48	0.906	0.918	0.932	0.927	0.914	0.891
NB	0.96	0.917	0.969	0.955	0.922	0.753
RC	0.905	0.924	0.904	0.984	0.928	0.938
RF	0.906	0.923	0.903	0.99	0.931	0.968
SVM	0.952	0.923	0.968	**0.991**	0.982	**0.991**

Table A2. Recall of five methods' recognizing positive emotion on JDN-gram, JDTSF, JD, JDResample, JDSmote and JDImmigration

Positive	JDN-gram	JDTSF	JD	JDResample	JDSmote	JDImmigration
J48	0.996	0.964	0.962	0.933	0.931	0.913
NB	0.893	0.955	0.949	0.938	0.969	0.99
RC	0.99	0.973	0.993	0.994	0.993	0.969
RF	0.996	0.979	**0.998**	0.996	0.996	0.978
SVM	0.979	0.969	0.978	0.979	0.978	0.969

Table A3. F1-measure of five methods' recognizing positive emotion on JDN-gram, JDTSF, JD, JDResample, JDSmote and JDImmigration

Positive	JDN-gram	JDTSF	JD	JDResample	JDSmote	JDImmigration
J48	0.949	0.941	0.947	0.93	0.922	0.902
NB	0.925	0.936	0.959	0.947	0.945	0.855
RC	0.946	0.948	0.946	0.989	0.959	0.953
RF	0.949	0.95	0.948	**0.993**	0.962	0.973
SVM	0.965	0.946	0.973	0.985	0.98	0.98

Table A4. Precision of five methods' recognizing negative emotion on JDN-gram, JDTSF, JD, JDResample, JDSmote and JDImmigration

Negative	JDN-gram	JDTSF	JD	JDResample	JDSmote	JDImmigration
J48	0.963	0.763	0.773	0.829	0.819	0.911
NB	0.602	0.717	0.768	0.852	0.91	0.985
RC	0.906	0.814	0.932	0.986	0.977	0.968
RF	0.963	0.851	0.98	**0.99**	0.989	0.977
SMO(SVM)	0.88	0.796	0.885	0.95	0.944	0.969

Table A5. Recall of five methods' recognizing negative emotion on JDN-gram, JDTSF, JD, JDResample, JDSmote and JDImmigration

Negative	JDN-gram	JDTSF	JD	JDResample	JDSmote	JDImmigration
J48	0.484	0.572	0.65	0.816	0.78	0.888
NB	0.813	0.569	0.85	0.891	0.794	0.675
RC	0.481	0.6	0.472	0.959	0.806	0.936
RF	0.484	0.591	0.463	0.975	0.814	0.968
SVM	0.753	0.597	0.841	0.978	0.956	**0.991**

Table A6. F1-measure of five methods' recognizing negative emotion on JDN-gram, JDTSF, JD, JDResample, JDSmote and JDImmigration

Negative	JDN-gram	JDTSF	JD	JDResample	JDSmote	JDImmigration
J48	0.644	0.654	0.706	0.822	0.799	0.899
NB	0.691	0.634	0.807	0.871	0.848	0.801
RC	0.629	0.691	0.627	0.972	0.884	0.952
RF	0.644	0.697	0.628	**0.983**	0.893	0.972
SVM	0.811	0.682	0.862	0.964	0.95	0.98

Table A7. Weighted average of Precision of five methods' recognizing emotions on JDN-gram, JDTSF, JD, JDResample, JDSmote and JDImmigration

Weighted Ave.	JDN-gram	JDTSF	JD	JDResample	JDSmote	JDImmigration
J48	0.916	0.892	0.906	0.899	0.887	0.901
NB	0.9	0.884	0.936	0.926	0.918	0.869
RC	0.905	0.906	0.909	0.984	0.942	0.953
RF	0.916	0.911	0.916	**0.99**	0.947	0.973
SVM	0.94	0.902	0.955	0.979	0.972	0.98

Table A8. Weighted average of Recall of five methods' recognizing emotions on JDN-gram, JDTSF, JD, JDResample, JDSmote and JDImmigration

Weighted Ave.	JDN-gram	JDTSF	JD	JDResample	JDSmote	JDImmigration
J48	0.911	0.899	0.91	0.899	0.888	0.901
NB	0.879	0.891	0.932	0.925	0.919	0.833
RC	0.905	0.91	0.906	0.984	0.939	0.953
RF	0.911	0.915	0.909	**0.99**	0.944	0.973
SVM	0.942	0.907	0.955	0.979	0.971	0.98

Table A9. Weighted average of F1-measure of five methods' recognizing emotions on JDN-gram, JDTSF, JD, JDResample, JDSmote and JDImmigration

Weighted Ave.	JDN-gram	JDTSF	JD	JDResample	JDSmote	JDImmigration
J48	0.898	0.893	0.907	0.899	0.887	0.901
NB	0.886	0.885	0.934	0.925	0.917	0.828
RC	0.893	0.905	0.893	0.984	0.937	0.952
RF	0.898	0.908	0.895	**0.99**	0.942	0.972
SVM	0.94	0.902	0.955	0.979	0.971	0.98

Table A10. Precision of five methods' recognizing positive emotion on JDN-gram, JD, JDResample, JDSmote and JDImmigration for evaluating the ability of generalization

Positive	JDN-gram	JD	JDResample	JDSmote	JDImmigration
J48	0.458	0.549	0.601	0.549	0.697
NB	0.686	0.752	0.782	0.566	0.448
RC	0.468	0.501	0.521	0.516	0.716
RF	0.463	0.498	0.534	0.541	0.693
SVM	0.639	0.733	0.726	0.737	**0.833**

Table A11. Recall of five methods' recognizing positive emotion on JDN-gram, JD, JDResample, JDSmote and JDImmigration for evaluating the ability of generalization

Positive	JDN-gram	JD	JDResample	JDSmote	JDImmigration
J48	0.971	0.893	0.856	0.856	0.881
NB	0.955	0.922	0.918	0.984	**0.996**
RC	0.975	0.979	0.984	0.984	0.934
RF	**0.996**	0.988	0.979	0.984	0.955
SVM	0.992	0.959	0.959	0.955	0.942

Table A12. F1-measure of five methods' recognizing positive emotion on JDN-gram, JD, JDResample, JDSmote and JDImmigration for evaluating the ability of generalization.

Positive	JDN-gram	JD	JDResample	JDSmote	JDImmigration
J48	0.623	0.68	0.706	0.669	0.778
NB	0.799	0.828	0.845	0.719	0.618
RC	0.633	0.663	0.681	0.677	0.811
RF	0.632	0.662	0.691	0.698	0.803
SVM	0.777	0.831	0.826	0.832	**0.884**

Table A13. Precision of five methods' recognizing negative emotion on JDN-gram, JD, JDResample, JDSmote and JDImmigration for evaluating the ability of generalization

Negative	JDN-gram	JD	JDResample	JDSmote	JDImmigration
J48	0.945	0.895	0.882	0.867	0.914
NB	0.964	0.945	0.944	0.982	0.99
Rcom	0.956	0.97	0.978	0.978	0.951
RF	**0.992**	0.981	0.975	0.98	0.964
SMO	**0.992**	0.969	0.969	0.966	0.962

Table A14. Recall of five methods' recognizing negative emotion on JDN-gram, JD, JDResample, JDSmote and JDImmigration for evaluating the ability of generalization

Negative	JDN-gram	JD	JDResample	JDSmote	JDImmigration
J48	0.303	0.555	0.655	0.573	0.768
NB	0.735	0.815	0.845	0.543	0.255
RC	0.328	0.408	0.45	0.44	0.775
RF	0.298	0.395	0.48	0.493	0.743
SVM	0.66	0.788	0.78	0.793	**0.885**

Table A15. F1-measure of five methods' recognizing negative emotion on JDN-gram, JD, JDResample, JDSmote and JDImmigration for evaluating the ability of generalization

Negative	JDN-gram	JD	JDResample	JDSmote	JDImmigration
J48	0.458	0.685	0.752	0.69	0.834
NB	0.834	0.875	0.892	0.699	0.406
RC	0.488	0.574	0.616	0.607	0.854
RF	0.458	0.563	0.643	0.656	0.839
SVM	0.793	0.869	0.864	0.871	**0.922**

Table A16. Weighted average of Precision of five methods' recognizing emotions on JDN-gram, JDTSF, JD, JDResample, JDSmote and JDImmigration for evaluating the ability of generalization

Weighted Ave.	JDN-gram	JD	JDResample	JDSmote	JDImmigration
J48	0.761	0.764	0.776	0.747	0.832
NB	0.859	0.872	0.883	0.825	0.785
RC	0.772	0.793	0.805	0.803	0.862
RF	0.792	0.799	0.808	0.814	0.862
SVM	0.859	0.88	0.877	0.88	**0.913**

Table A17. Weighted average of Recall of five methods' recognizing emotions on JDN-gram, JD, JDResample, JDSmote and JDImmigration for evaluating the ability of generalization

Weighted Ave.	JDN-gram	JD	JDResample	JDSmote	JDImmigration
J48	0.555	0.683	0.731	0.68	0.81
NB	0.818	0.855	0.872	0.709	0.535
RC	0.572	0.624	0.652	0.645	0.835
RF	0.561	0.619	0.669	0.678	0.823
SVM	0.785	0.852	0.848	0.854	**0.907**

Table A18. Weighted average of F1-measure of five methods' recognizing emotions on JDN-gram, JD, JDResample, JDSmote and JDImmigration for evaluating the ability of generalization

Weighted Ave.	JDN-gram	JD	JDResample	JDSmote	JDImmigration
J48	0.52	0.683	0.735	0.682	0.813
NB	0.821	0.857	0.874	0.706	0.486
RC	0.543	0.608	0.641	0.633	0.838
RF	0.524	0.601	0.661	0.672	0.825
SVM	0.787	0.854	0.85	0.856	**0.908**

- Proposed a topic sentence-based instance transfer method to process imbalanced Chinese product reviews
- Introduced a rule and supervision learning hybrid method for identifying topic sentence of a product review
- Incorporated feature set of the topic sentence to the feature space of sentiment classification
- Used a SMOTE-based method to overcome feature space inconsistency between source dataset and target dataset
- Result verified that our proposed methods helps SVM outperforms considering the ability of generalization