A topic sentence-based instance transfer method for imbalanced sentiment classification of Chinese product reviews

Tian, F, Wu, F, Chao, K-M, Zheng, Q, Shah, N, Lan, T & Yue, J

Original citation & hyperlink:
Tian, F, Wu, F, Chao, K-M, Zheng, Q, Shah, N, Lan, T & Yue, J 2016, 'A topic sentence-based instance transfer method for imbalanced sentiment classification of Chinese product reviews' Electronic Commerce Research and Applications, vol 16, no. March–April, pp. 66–76.
https://dx.doi.org/10.1016/j.elerap.2015.10.003

DOI 10.1016/j.elerap.2015.10.003
ISSN 1567-4223
ESSN 1873-7846

Publisher: Elsevier

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 University, 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;*

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 environment, the review-related data is imbalanced, and it leads to a problem in which minority class information tends to be ignored during 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 SMOTE (Synthetic Minority Over-sampling Technique) based method is adopted to overcome feature space inconsistency between 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 (http://www.jd.com) and Dangdang (http://www.dangdang.com) known as target dataset and source dataset respectively have been used in this paper. The experimental results verify that, considering the ability of generalization, our proposed method helps Support Vector Machine (SVM) to outperform J48, Naive Bayes, Random Forest and Random Committee when applied on N-gram, resample and SMOTE.

1 Introduction

In the last few years, we have witnessed a surge of interest in opinions mining automated systems for online product reviews [1-19]. The major supporting technology for opinion

mining systems includes topic modelling and sentiment analysis. Researchers and engineers/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 had 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, called as 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 training phase of classification model. The model trained from this kind of dataset that have low identification precision in minority classes has a phenomenon known as over-fitting of majority class. Some researchers employed sub-sampling strategy on imbalanced data to balance the class distribution of the dataset. This approach deteriorates the performance and generalization ability of the classification model trained on subsampled dataset. At the same time, different products (topics) from one data source may have imbalanced distribution characteristic in emotion class which could form different feature spaces with diverse data distribution in emotion classification. That is, the imbalanced distribution of emotion classes with different products (topics) represents different kind of interactions and mental state 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 [24] approaches. However, because these methods normally only rely on one dataset, the classification models that are trained on the dataset have over-fitting problem and lack the ability of generalization. For example, 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 data set of real-world for analysis, the classification performance is often degraded [25].

The methodologies behind classifiers that are trained on more than one auxiliary datasets have been widely adopted [26-30] in recent years in an attempt to address aforementioned problems of insufficient or homogeneous data by adopting the knowledge transfer learning method [31]. A simple method could directly combine an auxiliary dataset and an original dataset into a single dataset to train the classifier. As the tasks of emotion detection are strong domain/topic-dependent and the feature distribution of each topic has its own characteristics, we believe that such method will destroy the innate and unique features that exist in different domains and will decrease the recognition accuracy. In this paper, the task of topic sentence-based instance transfer 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. There are two datasets: target dataset (T) and source dataset (S) and dataset T can have different instance numbers in each class. The goal of instance transfer is as follow: assume that datasets S and T have the same classes of the sentiment analysis. In order to achieve the training task of sentiment classification model in T, it chooses the transferable instances of same class from S and transfer them to the corresponding class in T dataset 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'. Figure 1 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 similar number.

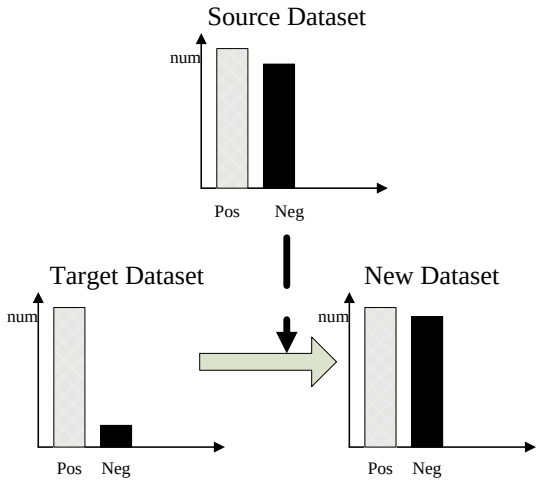

Figure 1. An instance transfer for imbalanced emotion classification

Challenges of implementing the above core idea depicted in Figure 1 are as follows: how to measure the transferability of instances in S and 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 (http://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 Unigram and Bigram of the product reviews corpora. In these two corpora, the products (topics) of Jingdong only include Laptop and PC, while the topics of Dangdang only includes digital product accessories. The number of items of Unigram and Bigram in the feature sets, JingDong and DangDang, are 1385 and 1258, and most of the items are different.

Inspired by the idea of 'topic' sentence [20] and [21] 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, of each product review. An algorithm for identifying 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

feature and frequency of emotion words and relevant nouns as shown in Table 1.

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 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 [33] 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.

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

Note that, the datasets we used in this paper contain two similar scales of minority emotion classes. The terms, sentiment and emotion are exchangeable, and there is no difference between them in this paper [22].

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 [1-9, 13, 15]. Chen *et al.* [3] use dependency parsing with shallow semantic analysis for Chinese opinion related expression extraction. Wu *et al.* [4] use phrase dependency parsing for opinion mining. Hu *et al.* [5] 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.* [13] adopted sentiment unigram and bigram as features. N-gram is also used as the features [13] [12]. Mukherjee *et al.* [7] abstract POS (part of speech)-tagged, all Nouns, direct neighbor and dependency relationship as the feature space of product feature. Cho *et al.* [15] 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.* [17] adopted HowNet lexicon for sentiment analysis of product reviews.

Currently, different kinds of data mining based techniques [1-19] are employed in sentiment analysis of product reviews. The researchers [2] applied text mining and NLP (natural language processing) 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, Android App Store. Mukherjee *et al.* [7] 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.* [10] 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 1000 hotel reviews from booking.com. Maks *et al.* [11] incorporated standard machine learning techniques naive Bayes and SVM into the domain of online Cantonese-written restaurant reviews to automatically classify user reviews as positive or negative. Kang *et al.* [13] proposed an improved Naïve Bayes algorithm for sentiment analysis of

restaurant reviews and got a higher accuracy than the original Naïve Bayes and SVM (support vector machine). Three supervised machine learning algorithms, Naïve Bayes, SVM and character based N-gram model are adopted for sentiment classification in [14]. Recently, Wang *et al.* [1] proposed a semi-supervised deep learning model which introduces supervised sentiment labels into traditional neural network language models for sentiment analysis. Both Fu *et al.* [17] and Bagheri *et al.* [18] adopted unsupervised methods for sentiment analysis of product reviews. After analyzing related literatures, we can conclude that most of the aforementioned methods are based on supervision approaches and only balanced datasets are used in their models.

Imbalanced data classification [25] 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 decrease in recognition performance [32]. The common methods to handle the above problem include data level sampling, cost sensitive learning, feature selection, feature weight adjustment and one-class learning[24][39].

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.* [31] and Barandela *et al.* [32] 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 proposed a boosting method based on cost sensitive training [34]. Zhou *et al.* proposed a method, which adopts neural network for cost sensitive learning to handle imbalanced problem [35].

The idea behind feature selection is to choose features, which are biased towards minority class in order to improve the learning outcome of minority class. Ogura *et al.* [24] proposed three metrics to select features, which are biased towards minority class, and they pointed out that these three metrics should be used synthetically. Liao *et al.* [36] proposed a method that selects features biased towards minority class by using feature distribution information. Wang *et al.* [1] 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 higher weight to features that is more important to minority class to solve imbalanced problem. Ying Liu *et al.* [37] proposed a method that adjusts features weight according to distribution ratio of minority class and majority class to increase the influence of minority class.

One-class learning is mainly applied to situations in which 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 investigated the limitation of a two class discrimination from the data with heavily unbalanced class proportions and pointed out that there is a consistent pattern of performance differences between one and two-class learning for all SVMs [38].

The research efforts mentioned above solve imbalanced problem aimed at a single target data set. These efforts take full use of the information of data itself to solve the problem. In recent years, researchers begin to adopt auxiliary datasets to solve the classification problem in different applications [26-31]. This paper 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 mentioned in Section 1, the challenge for the instance transfer method is how to measure the transferability of the instances (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, the on-line product reviews are a kind of paragraph-like writing-style. Inspired by the concept of 'topic' sentence used in automatic generation of abstract of literatures [20], we intend to apply the similarity of the topic sentences of two product reviews in the different data sets to measure the similarity of the two product reviews because 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 [22] is an object (such as products), event or domain, while the 'topic' sentence gives a strong indication of its overall subject [21].

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 paper, including the feature set of topic sentences, the features of product reviews have syntactic features, frequency features and N-gram features. Syntactic features and frequency features are shown in Table 1. In syntactic features, a Chinese sentiment lexicon base (include HowNet and our manually collected in our prior works [22], [23]) adopted. N-gram feature refers to the combinations of the words and has a strong dependency on data/corpus. In this paper, 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; 4) homogenizing incoherent feature spaces between transferred instances and dataset T to overcome issue of feature space inconsistency.

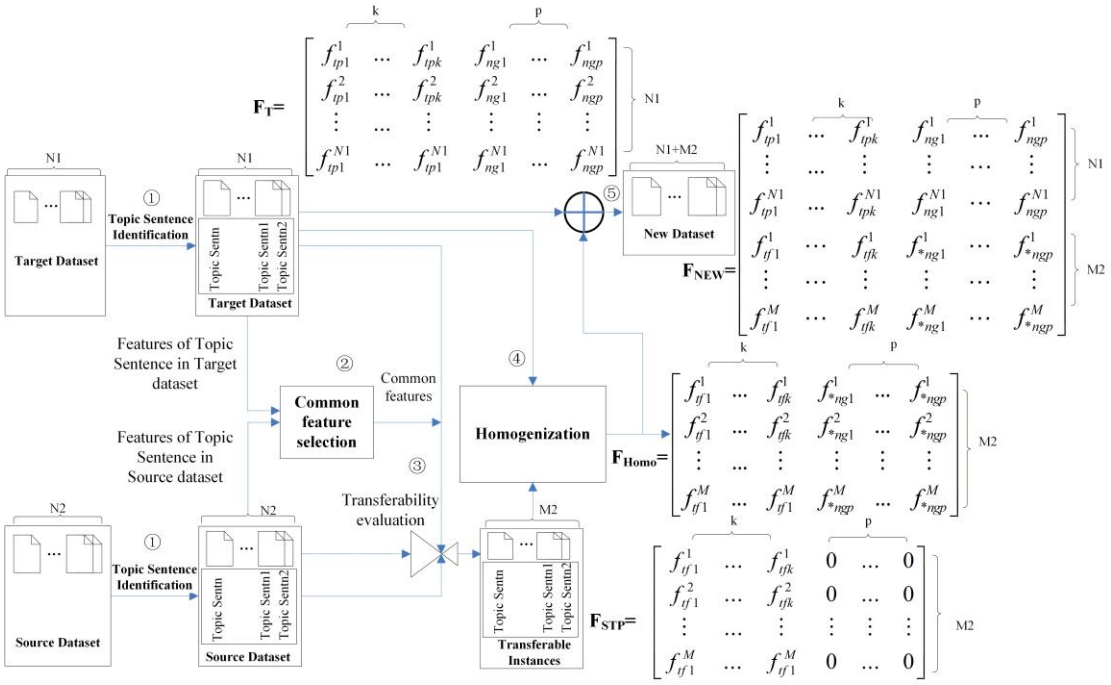

Figure 2. The frame diagram of our proposed approach

This paper proposes a new approach to solve the above problems. The frame diagram of the approach is shown in Figure 2. The dataset T contains N1 pieces of review instances and dataset S contains N2 pieces of review instances. F_T represents matrix of the feature values of dataset T and has k dimensions common features and p dimensions N-gram features. F_{STP} 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 0-value. $F_{H\text{omo}}$ 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 $F_{H\text{omo}}$. 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 in Figure 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 paper, common features is used to represent common features of topic sentences.

Step 3. **Transferability evaluation**. This step corresponds to the label ③ in Figure 2.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.

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.

Step 4: **Homogenization**. This step corresponds to the label ④ in Figure 2. It 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: **New dataset and Training**. This step corresponds to the label ⑤ in Figure 2. 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 details. Section 3.1 describes a topic-sentence identification method. Section 3.2 explains the method of selecting the common features of both T and S. Section 3.3 presents a cosine 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 Topic-sentence identification method

We have investigated the collected data and discovered that, if there exists 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 times the topic sentence of the product review located in first sentence or last sentence. So, a rule and supervision learning hybrid method for identifying topic sentence is proposed. In which, if there exists a title of a product review or has only one sentence in a product review, the method labels it as a topic sentence of the product review; otherwise, we use nouns (includes the product name, type and its producer), their frequency of occurrence in the review, relevant keywords of products, emotion words and their POS-tag, and their dependency in the first sentence and last sentence of each product review forms the feature set. We apply seven classification algorithms including J48, Random Forest, ADTree, AdaBoostM1, Bagging, Multilayer Perceptron and Naïve Bayes to the labeled datasets while using ten-fold cross-validation to test the performance of each classification model. According to the experimental results, we use Bagging method to identify that a topic sentence is either first or last sentence of a product review. This experiment and its results are shown in Section 4.2.

3.2 Common features selection in source and target datasets

In the feature set of topic sentence, category variables are majority variables. After computing, we found that the proportion of sum of the information gain [40] of common features between T and S has a relatively large proportion in both datasets. So, we have 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 (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 feature set to represent instances.

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:

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

$total = \mathrm{mode}(F_{S\cap T})$

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 $i = 1:M$

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

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

This process is shown in Figure 2, which is used for evaluating features by considering their weights in both datasets. The element position in two different ranked lists shows the difference of their importance in classification. There exists 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; $R_1 \neq R_2$; $F_{S\cap T}^{\mathrm{com}}$ represents the common features of S and T datasets, R is the dimension of $F_{S\cap T}^{\mathrm{com}}$. In computing process, Function *mode* calculates the element number in each dataset; If $total \neq \varnothing$ and 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 and 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 $index_{F_S}$. $F_{S\cap T}^{'} = \{f_{S\cap T}^{'1}, f_{S\cap T}^{'2}, f_{S\cap T}^{'3}, \ldots f_{S\cap T}^{'M}\} = \{f_{S\cap T}^{'1} \mid i = 1,2,\ldots M\}$,

$M = \mathrm{mod}e(F^{'}_{S\cap T})$ 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 the line 21, we get feature set $F^{\mathrm{com}}_{S\cap T}$.

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 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:

Step1: 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 [22] to normalize numerical features.

Step2: 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. Let $L = \{l_1, l_2, l_3, ..., l_N\} = \{l_p \mid p = 1,2,..., N\}$ denotes a set of class labels, N denotes the number of labels of classification tasks (in this paper, N=2, l_1 represents positive emotion, and l_2 represents negative emotion), and the formula of cosine similarity calculation is as follows:

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

Where, $InsTar^{l_P}(j))$ denotes an instance labeled with l_p in the target dataset; $j = 1,2,\cdots,n$ denotes 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; $i = 1,2,\ldots,K$ 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.

Step3: 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.

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 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. In this paper, we adopt 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

The above three sections provide detail 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. That is 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 following steps:

Step1: Collect experimental corpora: Two datasets were collected from two famous Chinese e-commerce portal, Jingdong (http://www.jd.com) and Dangdang (http://www.dangdang.com). These datasets are named as JingDong and DangDang, respectively. 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 [23] 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.

Step2: Identify the topic sentence of each review in Jingdong by employing the method described in Section 3.1 to evaluate the performance of the proposed method.

Step3: Based on JDTSF and DDTSF select 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, then determine the instances to be transferred from source dataset.

Step4: Carry out feature space homogenization processing method on the instances to be transferred according to the steps presented in Section 3.4.

Step5: Incorporate the transferred instances into each domain of target dataset according to the steps described in Section 3.4 and form a new training dataset JDImmigration. Note that, for comparison with traditional data sampling strategies/methods for imbalanced datasets, other two dataset, JDResample and JDSmote, are produced by applying resample and SMOTE to JD.

Step6: Apply five classification algorithms including J48, Random Forest, support vector machine, 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, in Figure 4-9 "RF" denotes Random Forest classification algorithm, Support Vector Machine [43] is denoted as "SVM", "RC" denotes Random Committee classification algorithm, and "NB" denotes Naive Bayes classification algorithm. The classification models, we take JDN-Gam, JD, JDResample, JDSmote, and CFImmigration as training set for the classification method and an extra dataset, JD634, in which 634 instances are collected from Jingdong as a testing dataset. To simplify the experiment, only J48, RF, RC, SVM, and NB are adopted in this step.

In the classification experiments, "P", "R" and "F" denote Precision, Recall and F1-measure respectively. Precision is 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. F1-measure is the harmonic mean of Precision and Recall. The classification experiments are carried out using Weka [44]. In addition, Weighted Average of each indicator in our experiment is the result of multiplying the value of the indicator in each emotion class (positive and negative) by corresponding weights and adding the total sum of the overall value, then dividing the total sum by total number of units.

4.2 Experimental results

After carrying out Step1 in Section 4.1, the number of features in the feature sets, of JD and DD, are 1418 and 1291, respectively. The numbers of N-gram in the two dataset are 1385 and 1258, respectively. The number of both positive instances and negative instances in DD is 2887. The number of positive instances in JD is 1600 while the number of negative instances in JD is 320.

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 2 shows weighted average of Precision, Recall and F1-measure of seven classification algorithms on 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 is 87.8%. The Bagging algorithm has shown the best performance.

The common features are selected by applying the method described in Section 3.2 and shown in Table 3.

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	twoFre
17	0.018511	degreeBelongComplement
18	0.016339	emotionVerb
19	0.014923	advBelongAtt
20	0.009263	degreeBelongAdverCount
21	0.005853	emotionNoun

After executing Step 4, the number of the transferred instances from DD is 1280 to make JDImmigration balanced. That is, the number of both positive and negative instances in JDImmigration is 1600.

In order to highlight the overall performance, we just list and analyze the weighted average of Precision, Recall and F1-measure in the following paragraphs. We collectively explain performance on negative and positive emotions at the end of this section and their experimental results are shown in corresponding tables in the Appendix.

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

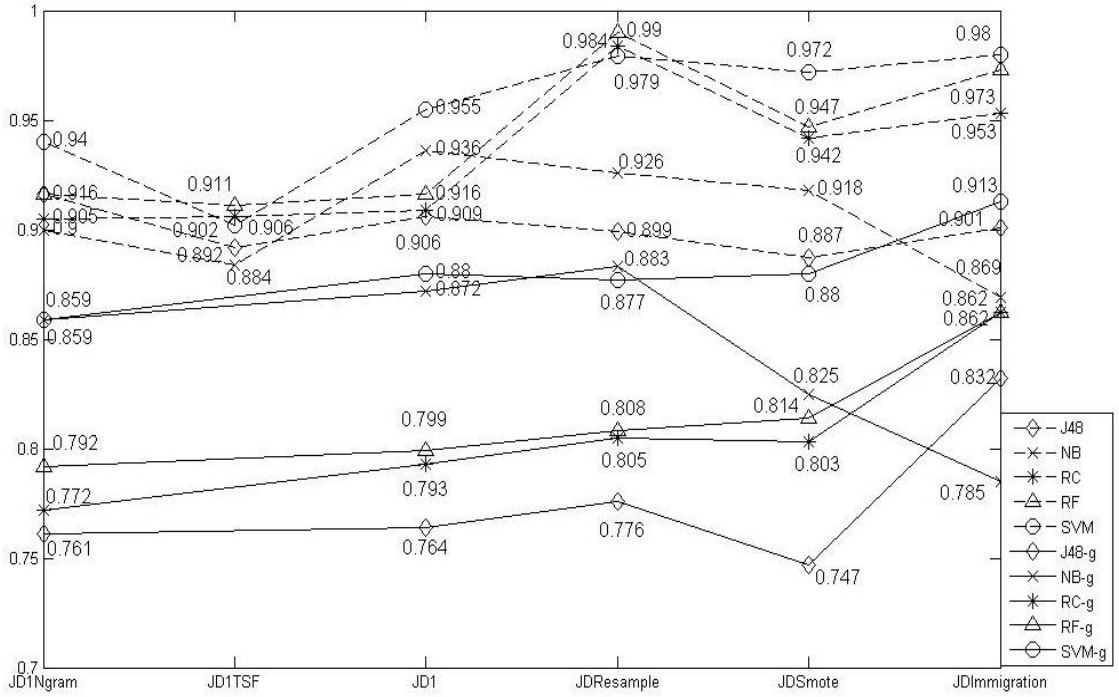

Figure 4. Weighted average of Precision of ten-fold crossing validation and the generalization ability evaluation

Figure 4 shows the weighted average of Precision of ten-fold crossing validation and the generalization ability evaluation. In Figure 4, dotted lines depict the results of ten-fold crossing validation when applying J48, NB, RC, RF and SVM on JDN-gram, JDTSF, JD, JDResample, JDSmote and JDImmigration, while 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.

According to five dotted lines shown in Figure 4, the best result of weighted average of Precision in ten-fold crossing validation is 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 applied on JDN-gram. Compared with the performance achieved on JDN-gram, the average improvement in weighted average of Precision of JD, JDResample, JDSmote and JDImmigration are 0.99%, 4.40%, 1.94% and 2.14%.

According to five solid lines shown in Figure 4, the best result of weighted average of Precision in generalization ability evaluation is achieved by applying SVM to JDImmigration, the value achieved is 0.913. Compared with the performance achieved on JDN-gram, the average performance improvement in weighted average of Precision of JD, JDResample, JDSmote and JDImmigrationare are 1.59%, 2.63%, 0.69% and 5.50% respectively. This show that our proposed method helps the adopted classification algorithms perform better than other methods in terms of the weighted average of Precision in generalization ability evaluation.

Note that, the percentage of average improvement is equal to the average of the difference of five method’s performance on JDN-gram and other dataset dividing by performance on JDN-gram. The percentages of average improvement mentioned in the following paragraphs are calculated in the same way.

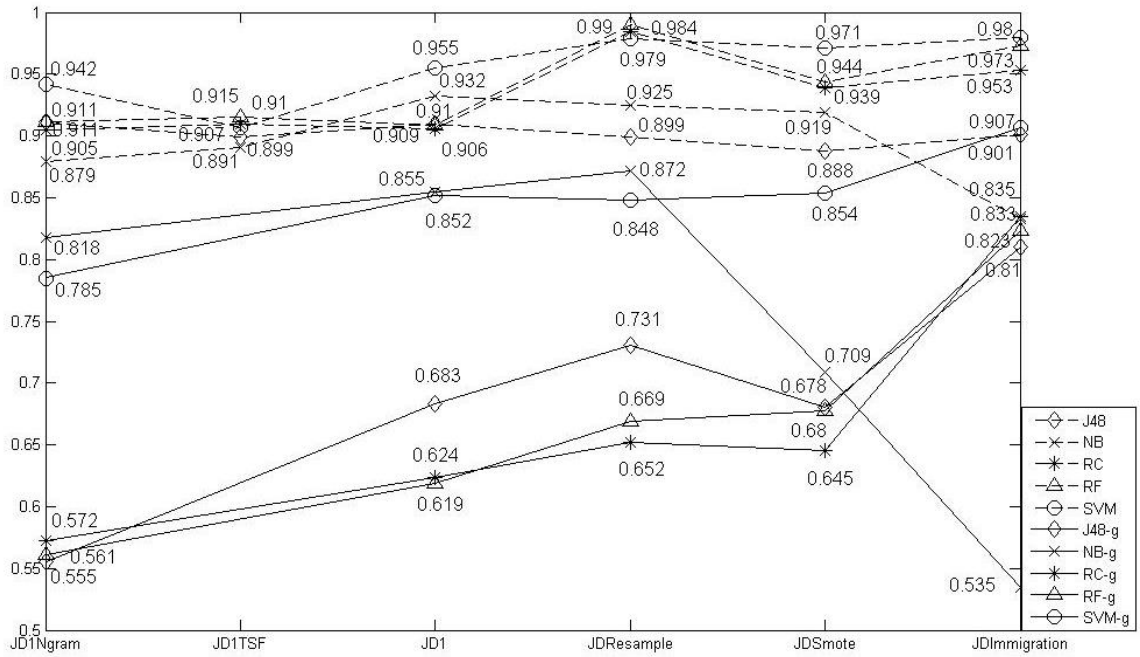

Figure 5. Weighted average of Recall of ten-fold crossing validation and the generalization ability evaluation

Figure 5 shows the weighted average of Recall of ten-fold crossing validation and the generalization ability evaluation. In the figure, dotted lines depict the results of ten-fold crossing validation when applying J48, NB, RC, RF and SVM on JDN-gram, JDTSF, JD, JDResample, JDSmote and JDImmigration, while 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 shown in Figure 5, the best result of weighted average of Recall in ten-fold crossing validation is achieved by applying RF to JDResample, which achieves value of 0.99. The four methods, NB, SVM, RF and RC applied on JDResample and JDSmote perform better than applied on JDN-gram. According to Table 5, compared with the performance achieved on JDN-gram, the average improvement in weighted average of Recall of JD, JDResample, JDSmote and JDImmigration are 1.44%, 5.05%, 2.50% and 1.96% respectively.

As shown by solid lines in Figure 5, the best result of weighted average of Recall in generalization ability evaluation is achieved by applying SVM to JDImmigration, a value of 0.907 is achieved. The four methods, J48, RC, RF and SVM applied on JD, JDResample, JDImmigration and JDSmote perform better than applied on JDN-gram. Compared with the performance on JDN-gram, the average performance improvement in weighted average of Recall of JD, JDResample, JDSmote and JDImmigration are 11.11%, 15.91%, 10.32% and 23.91%. This show that our proposed method helps the adopted classification algorithms perform much better than others in terms of the weighted average of Recall in generalization ability evaluation.

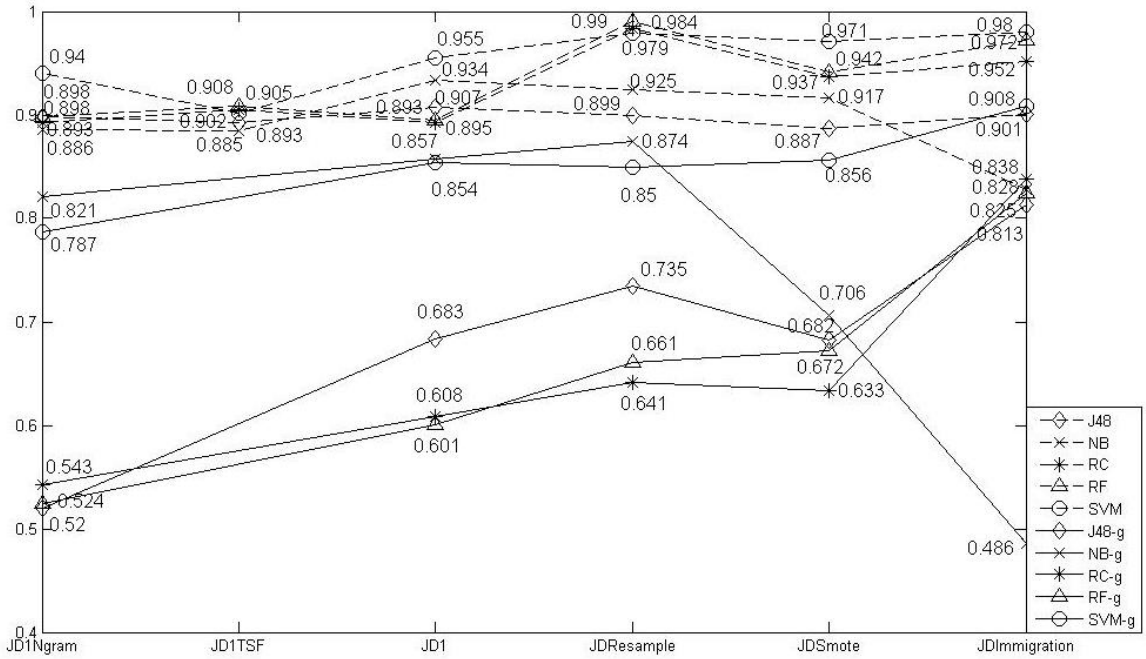

Figure 6. Weighted average of F1-measure of ten-fold crossing validation and the generalization ability evaluation

Figure 6 shows the weighted average of F1-measure of ten-fold crossing validation and the generalization ability evaluation. In the figure, dotted lines describe the results of ten-fold crossing validation when applying J48, NB, RC, RF and SVM on JDN-gram, JDTSF, JD, JDResample, JDSmote and JDImmigration, while 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 of weighted average of F1-measure in ten-fold crossing validation is achieved by applying RF to JDResample, and it achieves value of 0.99. The five methods J48, NB, SVM, RF and RC applied on JDResample perform better than applied on JDN-gram. Compared with the performance on JDN-gram, the average improvement in weighted average of F1-measure of JD, JDResample, JDSmote and JDImmigration are 1.54%, 5.82%, 3.08% and 2.58%.

According to five solid lines shown in Figure 6, the best result of weighted average of F1-measure in generalization ability evaluation is achieved by applying SVM to JDImmigration, the value of which is 0.908. The five methods applied on JD and JDResample perform better than applied on JDN-gram. Compared with the performance achieved on JDN-gram, the average improvement in weighted average of F1-measure of JD, JDResample, JDSmote and JDImmigration are 14.18%, 20.0%, 14.14% and 28.54%. This show that our method helps the adopted classification algorithm perform better than Resample and SMOTE in terms of the weighted average of F1-measure in generalization ability evaluation.

According to above experimental results, their analysis and Tables shown in Appendix, our conclusions are as follows:

- As it can be observed from dotted lines shown in Figure 4-6 and Tables A1-A9 that the performance of four classification methods (J48, RC, RF and SMV) when applied on the

feature set of topic sentences, JDTSF is good enough to achieve the performance comparable to JDN-gram considering the weighted average of Precision, Recall and F1-measure. This verifies that a 'topic' sentence is a strong indication of overall subject in each product review.

- In step 6 of the experiment, for ten-fold cross validation, JDResample has the best performance, which reached 0.99 on all three indicators, Precision, Recall and F1-measure. The outstanding performance of JDResample is mainly caused by over-fitting problem. This over-fitting problem is caused by applying resampling method to the minority class of JD, because the number of its negative instances is increased from 189 to 1880 by repeatedly sampling and the classification models trained on this kind dataset tend to recognize the information of features of duplicated instances in minority class of JD. Same thing happened to JDSmote.
- In order to evaluate the ability of generalization of Resampling, SMOTE and our proposed method, we conducted the step 7 of experiment. The results of step 7 show our proposed method has a stable improvement of Recall and F1-measure when applying J48, SVM, RC and RF. Moreover, F1-measure is the most common used method to comprehensively consider Precision and Recall indicators [45]. It effectively reflects the performance of the classification methods. Therefore, comparing the weighted average of F1-measure in the experiment, the average improvement of applying four classification methods (SVM, RF, RC and J48) to the immigration dataset produced by our proposed method outperforms other methods in the results of the generalization ability evaluation as well as the most results of ten-fold crossing validation. This verified that our proposed method overcome the influence of over-fitting problem and has an outstanding ability of generalization in terms of weighted average of F1-measure.
- Considering the weighted performance index, Precision, Recall and F1-measure, in the experiments for evaluating the ability of generalization, SVM on JDImmigration outperforms other classification methods as well as the data-level imbalanced data processing methods, resample and SMOTE.
- In our experiments, the improvement of performance of negative emotion is at a sacrifice of performance of positive emotion, this can be observed from Tables A1-A18. According to Tables A4-A6 and Tables A12-A15, the immigration dataset produced by our proposed method improves the classification performance on minority class (negative emotion) significantly in both ten-fold cross validation and generalization ability evaluation.

5 Conclusion

To effectively address the challenge of imbalanced sentiment analysis of product reviews, this paper 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 feature sets 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 [41], [42] for a real e-commerce platform. The service will analyze imbalanced sentiments in product reviews in real time.

Acknowledgement

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 Innovation Project of Shaanxi Province Key lab.

Reference

[1] Wang Y, Li Z, Liu , et al. Word Vector Modeling for Sentiment Analysis of Product Reviews [M]// Natural Language Processing and Chinese Computing. Springer Berlin Heidelberg, 2014: 168-180.

[2] Liu J, Sarkar M K, Chakraborty G. Feature-based Sentiment Analysis on Android App Reviews Using SAS® Text Miner and SAS® Sentiment Analysis Studio[C]// Proceedings of the SAS Global Forum 2013 Conference. 2013, 250.

[3] Chen M, Yao T. Combining Dependency Parsing with Shallow Semantic Analysis for Chinese Opinion-Element Relation Identification[C]// Universal Communication Symposium (IUCS), 2010 4th International. IEEE, 2010: 299-305.

[4] Wu Y, Zhang Q, Huang X, et al. Phrase Dependency Parsing for Opinion Mining[C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3. Association for Computational Linguistics, 2009: 1533-1541.

[5] Hu M, Liu B. Mining and summarizing customer reviews [C]// Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004: 168-177.

[6] Sharma R, Nigam S, Jain R. Mining of product reviews at aspect level [J]. International Journal in Foundations of Computer Science & Technology (IJFCST), 2014, 4(3): 87-95.

[7] Mukherjee S, Bhattacharyya P. Feature Specific Sentiment Analysis for Product Reviews [M]// Lecture Notes in Computer Science 7181 Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg, 2012: 475-487.

[8] Archak N, Ghose A, Ipeirotis P G. Show me the money! : deriving the pricing power of product features by mining consumer reviews[C]// In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '07). ACM, New York, NY, USA, 56-65.

[9] Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and trends in information retrieval [J]. Of Foundations & Trends in Information Retrieval Now Publ, 2008, 2(1-2):459-526.

[10] de Albornoz J C, Plaza L, Gervás P, et al. A Joint Model of Feature Mining and Sentiment Analysis for Product Review Rating [M] // Advances in information retrieval. Springer Berlin Heidelberg, 2011: 55-66.

[11] Maks I, Vossen P. Sentiment Analysis of Reviews: Should we analyze writer intentions or reader perceptions?[C]// Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria, 2013: 415–419.

[12] Zhang Z, Ye Q, Zhang Z, et al. Sentiment classification of Internet restaurant reviews written in Cantonese [J]. Expert Systems with Applications, 2011, 38(6): 7674-7682.

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

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

[15] Cho H, Kim S, Lee J, et al. Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews [J]. Knowledge-Based Systems, 2014, 71: 61–71.

[16] Zhang W, Xu H, Wan W. Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis [J]. Expert Systems with Applications, 2012, 39: 10283–10291.

[17] Fu X, Liu G, Guo Y, et al. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon [J]. Knowledge-Based Systems, 2013, 37: 186–195.

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

[19] Zhang L, Hua K, Wang H, et al. Sentiment Analysis on Reviews of Mobile Users [J]. The 11th International Conference on Mobile Systems and Pervasive Computing, Procedia Computer Science, 2014, 34: 458-465.

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

[21] Paice C D. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases[C]// In Proceedings of the 3rd annual ACM conference on Research and development in information retrieval (SIGIR '80). Butterworth & Co., Kent, UK, 1980:172-191.

[22] Tian F, Gao P, Li L, et al. Recognizing and regulating e-learners' emotions based on interactive Chinese texts in e-learning systems [J]. Knowledge-Based Systems, 2014, 55: 148-164.

[23] Tian F, Liang H, Li L, et al. Sentiment Classification in Turn-Level Interactive Chinese Texts of E-learning Applications[C]// 2012 IEEE 12th International Conference on Advanced Learning Technologies (ICALT). Rome, 2012: 480-484.

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

[25] He H, Ma Y. Imbalanced Learning-Foundations, Algorithms, and Applications[J].2013, IEEE Press

[26] Nguyen Q, Valizadegan H, Hauskrecht M. Learning classification with auxiliary probabilistic information[C]// Proceedings IEEE International Conference on Data Mining. IEEE International Conference on Data Mining 2011 (2011): 477–486.

[27] Tommasi T, Tuytelaars T, Caputo B. A testbed for cross-dataset analysis [J]. CoRR, 2014.

[28] Gong B, Sha F, Grauman K. Overcoming Dataset Bias: An Unsupervised Domain Adaptation Approach []. Big Data Meets Computer Vision: first international workshop on Large Scale Visual Recognition and Retrieval (BigVision) at NIPS, Lake Tahoe, NV, 2012.

[29] Heim E, Valizadegan H, Hauskrecht M. Relative Comparison Kernel Learning with Auxiliary Kernels [M]// Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science Volume 8724, 2014: 563-578.

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

[31] Pan S J, Yang Q. A Survey on Transfer Learning [J]. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2010, 22(10): 1345-1359.

[32] Barandela R, Valdovinos R M, Sánchez J S, et al. The imbalanced training sample problem: Under or over sampling? [M]. Structural, Syntactic, and Statistical Pattern Recognition. Berlin, Heidelberg: Springer, 2004: 806-814.

[33] Chawla N V. C4.5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure[C]// Proceedings of the ICML. Washington DC, 2003, 3.

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

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

[36] Liao Y, Pan X. Feature Selection Method on Imbalanced Text [J]. Journal of Xidian University, 2012, 41(4): 592-595.

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

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

[39] Satyam M, J A, Sanjeev S. Comparison of metrics for feature selection in imbalanced text classification [J]. Expert Systems with Applications, 2011, 38(5): 4978-4989.

[40] Han J, Kamber M. Data Mining: Concept and Techniques (Second Edition) [M]. The Morgan Kaufmann，2006.

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

[42] Huergo R S, Pires P F, Delicato F C, et al. A systematic survey of service identification methods [J]. Service Oriented Computing and Applications, 2014, 8 (3): 199–219.

[43] Platt J. Machines using sequential minimal optimization [J]. In Advances in Kernel Methods - Support Vector Learning: B, 1998.

[44] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. The WEKA Data Mining Software: An Update; SIGKDD Explorations, 2009, 11(1): 10-18.

[45] https://en.wikipedia.org/wiki/F1_score

Appendix

This appendix describes some experiment results haven't listed in main body of this paper.

Table A-1 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 A-2 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 A-3 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 A-4 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 A-5 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 A-6 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 A-7 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 A-8 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 A-9 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 A-10 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 A-11 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 A-12 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 A-13 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 A-14 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 A-15 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 A-16 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 A-17 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 A-18 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**