

# A Novel Cluster based Over-sampling Approach for Classifying Imbalanced Sentiment Data

Jing-Rong Chang, Long-Sheng Chen\*, Li-Wei Lin

**Abstract**—In social media, sentiments or online reviews are important information sources for product purchase decision-making. Usually, both favorable reviews and negative reviews from social media users may significantly impact companies' trade. Therefore, effective and efficient methods for sentiment classification have currently become the most concerning issues for companies. One of the best useful methods is machine learning. However, when employing these sentiment classification methods, the class imbalance problems which are caused by imbalanced data need to be considered since the performance accuracy of the majority class is often higher than that of the minority class. Therefore, this study aims to propose the Modified Cluster based over-Sampling (MCS) method which is expected to be a novel cluster based over-sampling method, for imbalanced sentiment classification. Some UCI data sets and three sentiment classification cases including two actual cases of imbalanced text comments regarding MP3 products and electronic commerce services were our research data which were employed to illustrate the usefulness of our proposed method. The experimental results indicate that our proposed method is superior to conventional re-sampling methods, such as over-sampling, cluster-based sampling, and decision tree algorithm.

**Index Terms**—Class imbalance problems, Sentiment classification, Cluster based sampling, Social media, Over-sampling, decision trees

## I. INTRODUCTION

SOCIAL media have been considered as the one of most important Internet applications (Chen et al., 2009; Chang et al., 2019). Recently, text content based social media, such as Facebook, micro-blog, Twitter, and Plurk have become crucial communication mechanisms used by an increasing amount of Internet users (Cohen and Krishnamurthy, 2006; Denecke and Nejdil, 2009; Singh et al., 2008; Chang et al., 2019; Chang et al., 2020). In the cyber world, textual comments, reviews, or opinions are usually powerful. These comments in social media can provide product information

and recommendations from the customer perspectives (Chang et al., 2020). Chang et al. (2019) also considers these textual reviews are important bases for product-purchase decision making of potential customers. Qiu et al. (2018) believes that online reviews are an important asset for users who are making business decisions. Consequently, a diversity of enterprises attempts to use social media as useful channels to promote their products and respond to customers. However, the results are not always positive. For instance, some negative comments related to companies' images or unfavorable evaluations of products will spread quickly and bring great damage to entrepreneurs. Besides, these negative product reviews can be treated as customer complaints which could provide useful information to improve specific enterprisers' services (Tian et al., 2016). Therefore, how to effectively classify online text comments' sentiments has become the most concerning problem for social commerce (Nakayama and Wan, 2019). Sentiment classification which divides sentiments into positive and negative can identify customers' emotions to help companies to carefully respond to customers' comments (Zhang et al., 2015).

Sentiment analysis has become important as the number of digital text resources increases with the development of information and communication technology (Ruz et al., 2020; Gokalp et al. 2020). The studies of sentiment classification have attracted not only researchers in 1990s, such as Kessler et al. (1997), Spertus (1997), and Argamon et al. (1998), but also the current researchers in the area of information management, data mining, and computer sciences, such as Chaovalit and Zhou (2005), Wiebe and Riloff (2005), Wilson et al. (2005), Li et al. (2018) and Liu et al. (2020). Specially, Li et al. (2020) even used news sentiments to predict stock prices.

Machine learning algorithms have been considered as one of the most effective solutions for successfully classifying social media users' sentiments (Zhang et al. 2015; Ibrahim and Abdulaziz, 2020; Xu et al., 2020). For example, Xu et al. (2020) proposed a continuous Naïve Bayes learning framework for large-scale and multi-domain e-commerce platform product review sentiment classification. Alam et al. (2020) used convolutional neural networks (CNN) to analyze sentiments in social media. Asif et al. (2020) used Support Vector Machines (SVM) to classify the sentiments of extremism and reported that SVM was superior to Naïve Bayes (NB).

However, Liu et al. (2009) indicated that the natural distribution of textual data used in text classification is often imbalanced. Such imbalanced data, for example, huge amounts of favorable comments overwhelm negative comments and vice versa, often cause class imbalance

Manuscript received August 01, 2021; revised October 15, 2021. This work was supported in part by the Ministry of Science and Technology, Taiwan (R.O.C.) under Grant MOST 110-2410-H-324-003, and Chaoyang University of Technology under Grant 110F0021109.

Jing-Rong Chang is an Associate Professor of Department of Information Management, Chaoyang University of Technology, Taichung 413310, Taiwan (R.O.C.) (e-mail: chrischang@cyut.edu.tw).

Long-Sheng Chen is a Professor of Department of Information Management, Chaoyang University of Technology, Taichung 413310, Taiwan (R.O.C.) (corresponding author's phone: +886-4-23742302; fax: +886-4-23742303; e-mail: lschen@cyut.edu.tw).

Li-Wei Lin is a master student of Department of Information Management, Chaoyang University of Technology, Taichung 413310, Taiwan (R.O.C.) (e-mail: s9714626@gmail.com).

problems when applying machine learning methods to sentiment classification of textual data (Chen et al., 2008; Liu et al., 2008; Suresh et al., 2008; Su and Hsiao, 2007; Zhang et al., 2015). Therefore, class imbalance problems have been majorly concerned and currently attracted many researchers in various fields since the classification performance accuracy of the majority class is often higher than that of the minority class (Chen et al., 2008; Vashishtha and Susan, 2019; Fu et al., 2020). Particularly, class imbalanced problems were considerable issues in text mining domain, such as automatic text categorization (Liu et al., 2009; Ogura et al., 2011; Zheng et al., 2004; Chen et al., 2010; Sun et al., 2009), disease detection (Richhariya and Tanveer, 2020), fault diagnosis (Qian and Li, 2020), author identification (Stamatatos, 2008), direct marketing and churn prediction (Wong et al., 2020), credit risk prediction (Zanin, 2020), political and religious extremism detection (Asif et al., 2020), and spam detection (Sakkis et al., 2001). Specially, in the field of political detection, Chau and Xu (2007) mentioned that negative comments related to anti-Blacks, suicide, bomber making, drugs, and porn were comparatively much fewer than normal opinions. Asif et al. (2020) added some political, religious, or social issues which would cause extremism among people were depicted by their sentiments on social media. However, the existing researchers agreed that identifying minor comments effectively is more important than detecting normal comments in many textual classification cases. For instance, for spam identification, Cohen and Krishnamurthy (2006) concluded that although the amount of regular emails was much less than the amount of spam emails, identifying regular emails was more important than detecting spam ones. Stamatatos (2008) focused on author identification. The text categorization problems found in this study were that extremely few training texts were collected from some selective authors or a significant variation in the text-length was present among the available training texts of the selective authors.

Recently, many researchers have studied imbalanced textual data (Wang et al., 2013; Tian et al., 2016; Li et al., 2018). For example, Li et al. (2018) proposed a domain-adaptive model that incorporates universal and domain-specific knowledge for imbalanced text sentiment classification. Wang et al. (2013) presented a sample cutting method for imbalanced text sentiment classification. Tian et al. (2016) proposed a topic sentence-based instance transfer method for imbalanced sentiment classification of Chinese product reviews. Liu et al. (2009) presented a simple probability-based term weighting scheme in which two critical information ratios were used to identify documents in minor categories. Stamatatos (2008) attempted to solve imbalanced textual data by segmenting the training texts into text samples basing on the size of the class, then produced a fairer classification model. Padurariu and Breaban (2019) employed a cost-sensitive method for imbalanced text data. Zheng et al. (2004) investigated the effectiveness of several feature metrics, such as information gain, Chi-square, correlation coefficient, and odds ratios based on Naïve Bayes and SVM for imbalanced text categorization classification. In the work of Xi et al. (2019), they considered that least squares support vector machine (LSSVM) was an effective method

for solving class imbalance problems. Chen et al. (2011) proposed a new method to improve text categorization under class imbalance by exploiting the semantic context in text documents. Sun et al. (2009) used SVM to conduct a comparative study on the effectiveness of these strategies in the context of imbalanced text classification. El-Alfy and Al-Azani (2017) compared different classifiers for polarity determination in highly imbalanced short Arabic text datasets with and without the Synthetic Minority Over-Sampling Technique (SMOTE).

Following this trend, this study aims to propose a Modified Cluster-Based Over-Sampling (MCS) method to tackle the class imbalance problems in textual sentiment classification. Three cases including two actual cases of imbalanced textual comments regarding MP3 products and electronic commerce services were our research data which were employed to illustrate the effectiveness of our proposed MCS method. The experimental results were compared with traditional techniques, such as over-sampling, under-sampling, and cluster-based sampling.

## II. RELATED WORKS

In recent years, imbalance data classification which is an intrinsic characteristic of multi-label data (Liu and Tsoumakas, 2020) and class imbalance problems have been so important that they have attracted a wide diversity of researchers in data mining areas (Andrzej, 2019; Henriquez et al., 2019; Hendry and Chen, 2019; Richhariya and Tanveer, 2020; Bria et al., 2020). Therefore, the amount of published papers regarding the class imbalance problems has been continuously increasing in IEEE and ACM (He and Garcia, 2009). In related works, several methods have been proposed to solve class imbalance problems. These methods can be divided into two groups. Methods in the first group are algorithm-oriented approaches which aim to propose new learning algorithms or modify existing methods. This group includes various approaches, such as Support Vector Machines (SVM) (Richhariya and Tanveer, 2020), one class learning, neural networks (NN), and information granulation-based methods. Particularly, Wu and Chang (2005) used SVM to tackle class imbalance problems while Richhariya and Tanveer (2020) presented a reduced universum twin SVM for class imbalance learning. Moraes et al. (2013) applied both SVM and Artificial Neural Networks (ANN) to classify a textual review. Yan et al. (2009) proposed an adjustment method of the separating hyperplane in SVM for the credit access application in commercial banks. Orriols-Puig et al. (2009) aimed to study the behavior of Michigan-style learning classifier systems on imbalanced domains and to use the lessons provided by the analysis to improve the modeling of rare classes. Manevitz and Yousef (2002) used one class learning SVM to detect rare events.

In addition, deep learning, logistic regression, and neural networks have been employed to tackle this issue. For instance, Bria et al. (2020) proposed a two-stage deep learning framework to deal with the high-class imbalance encountered during training of small lesion detectors medical images. Williams et al. (2009) proposed a modified logistic regression method for imbalanced remote-sensing classification problems. Qian and Li (2020) proposed a class imbalance-robust network which combines feature extraction

and classification for bearing fault diagnosis.

The second group includes several re-sampling techniques. García et al. (2019) suggested re-sampling methods, either by under-sampling the majority class or by over-sampling the minority class, and then the most powerful techniques were selected for solving the class imbalance problems. The re-sampling techniques include (1) under-sampling, (2) over-sampling, (3) cluster-based sampling and kernel density estimation-based sampling, (4) cost sensitive, and (5) Mahalanobis Distance (MD) based two phase learning. In the under-sampling methods, the minority population is kept intact, while the majority population is under-sampled (Liu and Tsoumakas, 2020; Hamidzadeh et al., 2020). In contrast in the over-sampling methods, the minority examples are over-sampled so that the desired and fair class distribution is obtained in the training set (El-Alfy and Al-Azani, 2017; García et al., 2019; Raghuwanshi and Shukla, 2020). Differently in the cluster-based sampling methods proposed by Altincay and Ergun (2004), Chen et al. (2009b), and Ofek et al. (2017) and the kernel density estimation-based sampling methods suggested by Kamalov (2020), the representative examples are randomly sampled from clusters or kernel density estimation. Furthermore, in the cost sensitive methods, the prediction accuracy is improved by adjusting the cost (weight) for each class (Min and Liu, 2009; Weiss and Provost, 2003; Zhang et al., 2019; Wong et al., 2020). Finally, in the Mahalanobis Distance (MD) based two phase learning methods, MD is firstly used to screen the majority examples which are ensured 100% that they are truly majority (Chen et al., 2009a).

Moreover, Wong et al. (2020) proposed cost-sensitive deep neural networks (NN) and cost-sensitive deep NN ensemble methods to address class imbalance problems. Ofek et al. (2017) proposed a fast clustering-based under-sampling method for classifying imbalance data. Liu et al. (2008) presented a weighted rough sets model in which weighted information entropy was used to deal with imbalanced data. Suresh et al. (2008) proposed two risk-sensitive loss functions to solve the class imbalance problems. Fernández et al. (2010) attempted to improve the behavior of fuzzy rule-based classification systems in imbalanced data-sets by adapting the 2-tuples based genetic tuning approach.

Generally, although the methods in the first group may have better performance than the methods in the second group, the methods in the second group are simpler and long training time are unnecessary. Therefore, this study focuses on re-sampling techniques which are to modify the class distributions of the training data by different sampling strategies. Over-sampling and under-sampling are the simplest methods. However, over-sampling cannot gain new information about the minority class, since under-sampling may lose useful information about the majority class (Japkowicz and Stephen, 2002). These supervised methods lack a rigorous and systematic treatment on imbalanced data and they still have some drawbacks (Huang et al., 2004). To enhance over-sampling and under-sampling, cluster-based sampling techniques (Altincay and Ergun, 2004) have been proposed. However, the number of clusters and how to choose representative examples are difficult to be determined in cluster-based sampling methods (Hulse et al., 2009).

Therefore, a cluster-based over-sampling approach has been presented to select those who have nearest distances from centroids of the built clusters to be representative examples for learning (Jo and Japkowicz, 2004; Ofek et al., 2017). Cohen et al. (2006) used the K-mean algorithm to cluster examples and then employed the central point of a constructed cluster to represent the whole cluster. Li et al. (2008) presented an under-sampling method based on variable self-organizing map (SOM) clustering. They firstly used SOM to cluster data and then deleted some examples in the clusters that have high purities to improve the imbalanced situation.

Additionally, some researchers indicated that data complexity or data structure, such as over-lapping, lack of representative data (He and Garcia, 2009), and small disjuncts (Weiss and Provost, 2003) might be the causes of the class imbalance problems. Also, this work aims to study the sentiment classification problems; therefore, textual data which has some unique characteristics, such as high dimensionality and small sample size (i.e. the number of samples might be smaller than the number of attributes) will need to deal with because characteristics of textual data may influence the performance of classifiers.

To sum up, this study aims to propose a Modified Cluster based Sampling (MCS) method which introduces two objective indexes, purity and entropy, to determine the number of clusters. In addition, we implemented 2 sampling strategies to select representative examples to find a robust cluster-based sampling method to tackle the class imbalance problem in social media users' sentiment classification. Several data sets from UCI data bank and three real cases of product reviews were experimented to illustrate the effectiveness of our proposed MCS method. Also, some experiments were implemented to evaluate the influences of handling textual data including both high dimensionality and small sample size.

### III. PROPOSED METHODOLOGY

#### A. *Implemental Procedure*

This section will introduce the detailed implemental procedure of the proposed MCS approach which can be separated into three phases. As shown in figure 1, they are "clustering", "selecting representative examples from constructed clusters", and "learning".

In the clustering phase, we gathered similar examples into clusters. The objective of Phase II is to select representative examples from the constructed clusters. In this phase, we used over-sampling techniques which mean duplicating the minority examples until the class imbalanced situation was improved. Consequently, for selecting representative examples, the majority examples were kept intact and the minority examples were duplicated by utilizing two strategies, named MCS1 and MCS2. Depending on measuring the Euclidean distances between examples and the central point of the built cluster, MCS1 selects the representative minority examples and then employs over-sampling technique to duplicate these minority examples until the class imbalance situation was improved. MCS2 uses three data characteristics (minimum, median, maximum) of every single one cluster as

the representative minority examples. Then those selected representative minority examples were joined together with the majority examples to be the training set. In the final phase, the objectives of MCS1 and MCS2 were to build classifiers from this balanced training set. In this study, we employed a decision tree (C4.5) to be our learner. Moreover, we compared our proposed methods with traditional methods under consideration of different dimensionality and small sample size of data. Hopefully, we could find one robust cluster-based sampling method for imbalanced sentiment classification.

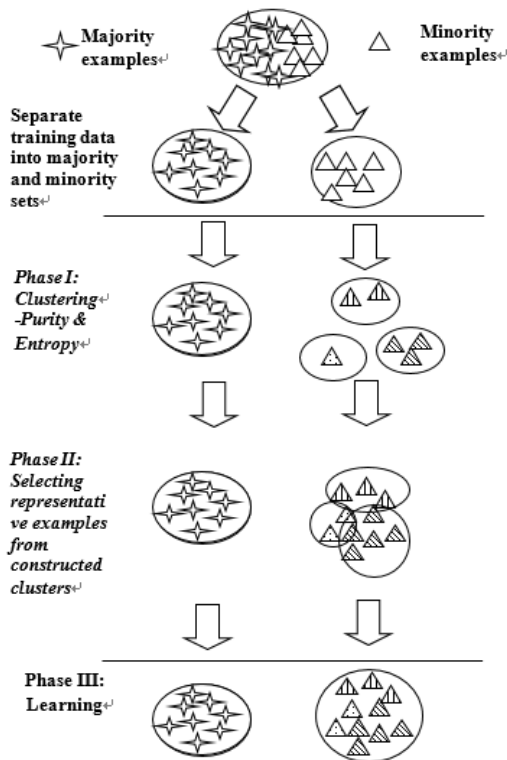


Fig. 1. The concept of the proposed MCS1 method

### B. MCS1

This section will introduce the algorithm of MCS1. In MCS1, we introduce two indexes, purity and entropy whose definitions can be found in equations (1) & (2), to determine the suitable number of clusters.

$$Entropy = \sum_{r=1}^k \frac{n_r}{n} E(C_r) \quad (1)$$

$$Purity = \sum_{r=1}^k \frac{n_r}{n} P(C_r) \quad (2)$$

Where,

$$E(C_r) = -\frac{1}{\log q} \sum_{i=1}^q \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r} \quad (3)$$

$$P(C_r) = -\frac{1}{n_r} \max_i (n_r^i) \quad (4)$$

and  $q$  denotes the number of class labels,  $n_r^i$  means the percentage of class label  $r$  in cluster  $i$ ,  $k$  is number of constructed clusters.

Unlike traditional over-sampling method which merely randomly duplicates the minority examples until the

imbalanced situation is improved, after clustering, MCS1 firstly selects the minority examples which are close to the central point of built clusters to be our representatives of each constructed cluster. Then, we used oversampling technique to duplicate the representative minority examples. Next, these duplicated minority examples were combined together with the majority examples to be the training set. Finally, we built a classifier from this balanced training set. The detailed steps for MCS1 are as follows:

#### Phase I: Build Cluster

Step 1: Construct clusters from the minority objects.

K-mean has been utilized to construct cluster objects. Purity and entropy indexes are also used to determine number of clusters.

Step 2: After clustering, we separate majority and minority examples into two groups. We do nothing for the majority examples. We only apply over-sampling technique to the minority examples.

#### Phase II: Cluster based sampling

Step 3: Compute the Euclidean distance between objects and the central point of cluster.

Step 4: Select representative samples which are close to the central point. Those objects will be the representatives of the constructed clusters.

Step 5: Implement over-sampling technique. The selected minority examples will be duplicated till their amount is equal to the majority examples.

#### Phase III: Build Classifiers

Step 6: Combine those duplicated minority examples (in Step 5) with the original majority examples to be the training data set.

Step 7: Build a classifier by using decision tree (C4.5).

### C. MCS2

MCS2 uses another kind of over-sampling method to produce representative minority set in phase II. Unlike MCS1, MCS2 uses three data characteristics such as minimum, median, and maximum to be the represent one constructed minority clusters. In other words, we use three statistically descriptive values to denote every single one cluster. Detailed algorithm of implementing MCS2 can be found as bellow.

#### Phase I: Build Cluster

Step 1: Construct clusters from the majority examples.

In this work, K-mean will be utilized to cluster objects. But, we don't implement purity and entropy to determine the number of clusters. According to the amount of majority examples, we determine the suitable number of clusters.

Step 2: For each cluster, we separate majority and minority examples into two groups. The majority population is kept intact.

#### Phase II: Cluster based sampling

Step 3: Compute values of minimum, median, and maximum for every minority class cluster. These three values will become the representatives of the constructed clusters of minority examples.

#### Phase III: Build Classifiers

Step 4: Combine those representatives with the original

majority examples to be the training data set.

Step 5: Build a classifier. In this study, we use decision tree (C4.5) to construct classifiers.

D. Traditional Re-sampling Methods for Imbalanced Data

In this work, the proposed MCS1 and MCS2 will be compared with traditional re-sampling techniques for solving class imbalance problem, such as Li’s method, cluster based under-sampling (CBS), and over-sampling method. Let’s briefly introduce these comparative methods.

Firstly, we introduce Li’s method (2008). After clustering the majority examples, this method uses the under-sampling strategy of removing the examples whose purities is higher than others to reduce the amount of majority examples. However, there is slightly different from the original approach. For example, we only conduct the single one class examples and use two indexes to determine the number of clusters. The detailed algorithm of Li’s method can be found as follows.

Phase I: Build Cluster

Step 1: Construct clusters from the majority objects.

In this step, K-mean and two indexes, purity and entropy, have been employed to construct clusters.

Step 2: Separate majority and minority examples into two groups. The minority population is kept intact.

Phase II: Cluster based Sampling

Step 3: Rank the purities of all constructed clusters.

Step 4: Remove those examples in the clusters whose purities are higher than others till the imbalanced situation has been improved.

Phase III: Build Classifiers

Step 5: Join those representative majority examples and the original minority examples together to be the training data set.

Step 6: Build a classifier.

In CBS method, we use two indexes to determine the suitable number of clusters. Then we use under-sampling method to select representative examples from both majority and minority examples. Over-sampling method is to reduplicate the minority examples till the imbalanced situation has been improved. DT means we do nothing for imbalanced data and merely implement decision tree algorithm.

E. Measurement Index

This section will introduce some indexes used in this study. First one is overall accuracy (OA). But, OA is not enough for imbalanced sentiment classification. OA considers misclassification errors to be equal. But as we know, a highly imbalanced class situation does not have equal costs that favor the minority class, which is often the class of primary interest. Consequently, in addition to OA, we use positive accuracy (PA), negative accuracy (NA), and geometric mean of PA and NA (G-Mean) to evaluate classifiers. The definition of OA is as equation (5).

$$\text{Overall Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

, where TP is the number of true positive; TN denotes the number of false negative;

FP means the number of false positive; TN represents the number of true negative.

In this study, PA and NA represent the ability of detecting the positive (majority) and negative (minority) examples, respectively. PA and NA are defined as

$$PA = \frac{TP}{TP + FN} \quad (6)$$

$$NA = \frac{TN}{FP + TN} \quad (7)$$

Another index G-mean defined in equation (8) has been introduced. This index is to maximize the accuracy on each of two classes while keeping these accuracies balanced. For instance, a high PA by a low NA will result in a poor G-mean.

$$G - \text{mean} = \sqrt{PA \times NA} \quad (8)$$

IV. IMPLEMENTATION

A. Employed Data

In this study, we firstly use five data sets from UCI machine learning repository which are available from the website (www.ics.uci.edu/~mllearn) to evaluate the effectiveness of our proposed MCS method in general case. In order to create imbalanced data set, we combine all classes together except the minority class. Therefore, we transfer multi-class into binary-class classification problem. Table 1 summarizes the basic information of the five used data sets.

TABLE I  
THE EMPLOYED DATA SETS

Data set	Data size	No. of attributes	Attribute value	Class distribution
Haberman	306	4	Integer	Survived: 74% Died*: 26%
Ionosphere	351	34	Integer, Real	Good: 69% Bad*: 31%
TAE	151	5	Categorical, Integer	Low*: 32.5% Med: 33.1% High: 34.4%
Diabetes	759	9	Integer	Healthy: 66% Diabetic*: 34%
Contraceptive Method Choice (CMC)	1473	9	Categorical, Integer	No-use: 43% Long-term*: 22% Short-term: 35%

Note: In class distribution column, “\*” indicates the minority class, when we combine other class examples to be our majority class.

TABLE II  
THE EMPLOYED TEXTUAL DATA SETS

No	Data Set	Notation	No. of attributes	Data Size	Class distribution
1	MP3 product evaluation	MP3	300	349	Positive: 305 Negative: 44
2	EC service comment	EC	66	200	Positive: 140 Negative: 60
3	Movie Review	Movie	325	1000	Positive: 700 Negative: 300

Besides, this work focuses on imbalanced sentiment classification problem. Therefore, we employs three imbalanced sentiment data sets including two real cases from

real world social media users' comments and one famous movie reviews database (available at <http://www.cs.cornell.edu/people/pabo/movie-review-data/>). Table 2 summarizes the brief background of these three employed imbalanced sentiment data sets. The first two data sets are from "reviewcenter ([www.reviewcentre.com](http://www.reviewcentre.com))". By focusing the topics of "electronic commerce service (EC)" and "MP3 product evaluation (MP3)", we collect 200 and 349 reviews, respectively. There are 140 positive and 60 negative comments in EC data set and the amount of attributes is 66. In MP3 data set, we collect 304 positive and 44 negative comments, and there are 300 attributes to describe MP3 data. Moreover, because these comments have no class labels, we use the 10-star rating system in "reviewcenter" website to define users' sentiments. For example, a comment will be labeled as positive (negative) if the rate is above 7 stars (below 4 stars). Those comments whose rates are between 4 and 7 stars have been disregarded. The last data set comes from movie reviews database. In order to create imbalanced data set, we randomly select 700 positive and 300 negative comments to be our experimental corpus. The dimensionality of movie review data is 325. Moreover, four-fold cross validation experiment has been implemented in this study.

In addition, the shareware Rubryx (<http://www.sowsoft.com/rubryx>) has been utilized to extract key words (attributes of textual data) in this work. Rubryx segments words based on n-gram (unigram, bigrams, and tri-grams) features. Before extracting n-gram key words, some frequently used stop words should be removed. Readers can find a useful stop word list which is available at [http://www.dcs.gla.ac.uk/idom/ir\\_resources/linguistic\\_utils/s\\_top\\_words](http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/s_top_words). Each comment is converted into a vector of terms (keywords) with term frequency.

**B. Results of Five UCI Data Sets**

This section will evaluate the effectiveness of the proposed MCS1 and MCS2 in UCI data sets. Table 3 summarizes the implemental results of our proposed MCS1 and MCS2 and traditional methods on five UCI data sets. Even implementing 4 fold cross validation experiment, we merely list mean value and standard deviation (SD) for the purpose of being easily understood. In this table, "DT" means we do nothing for imbalanced data and just implement decision tree algorithm. PA and NA denote the ability of identifying the majority examples and minority examples, respectively. In the results of DT, we found PAs are very high and NAs are very low in "Haberman", "TAE", "CMC", and "Diabetes" data sets. That means there are serious class imbalance problems in these four data sets. But, in "Ionosphere" set, this problem is not very bad.

**C. Results of Five UCI Data Sets**

This section will evaluate the effectiveness of the proposed MCS1 and MCS2 in UCI data sets. Table 3 summarizes the implemental results of our proposed MCS1 and MCS2 and traditional methods on five UCI data sets. Even implementing 4 fold cross validation experiment, we merely list mean value and standard deviation (SD) for the purpose of being easily understood. In this table, "DT" means we do nothing for

imbalanced data and just implement decision tree algorithm. PA and NA denote the ability of identifying the majority examples and minority examples, respectively. In the results of DT, we found PAs are very high and NAs are very low in "Haberman", "TAE", "CMC", and "Diabetes" data sets. That means there are serious class imbalance problems in these four data sets. But, in "Ionosphere" set, this problem is not very bad.

Since G-mean can take both PA and NA into consideration, we firstly use this index to measure the performances of proposed methods. From table 3, all G-means of five data sets indicate that MCS2 has the best performance compared than MCS1, CBS, Li's method, over-sampling, and DT. This situation also could be easily confirmed from Figure 2. From this figure, readers can easily find the peaks of performance curves locate at MCS2.

Moreover, MCS2 also can increase over accuracy in most experiments such as "TAE", "CMC", "Ionosphere" and "Diabetes". MCS1 is only slightly better than other traditional re-sampling methods in "TAE", "Ionosphere", and "Diabetes." MCS1 is worse than CBS in "Haberman" and "CMC" data. Therefore, from the results, we can conclude that MCS2 is more effective than MCS1. In addition, compared with traditional re-sampling techniques, MCS2 can not only increase G-mean, but also increase the overall accuracy.

TABLE III  
THE RESULTS OF FIVE UCI DATA SETS

Index	Proposed Methods				Traditional Methods							
	MCS1 (%)		MCS2 (%)		CBS (%)		Li's method (%)		Over-sampling (%)		DT (%)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Haberman												
PA	60.3	11.5	63.3	16.6	72	134	42.4	10.9	594	152	93.8	8.5
NA	65.8	11.9	69.1	15.2	232	455	79.36	15.7	715	203	18.9	22.5
OA	61.5	6.6	64.4	9.8	68	54	51.59	5.4	62	98	75.4	3.5
GM	62.2	5.2	<b>64.9</b>	<b>7.7</b>	634	57	57.08	3.20	637	105	28.5	33.1
TAE												
PA	73.8	13.0	81.8	5.0	625	196	38.4	10.5	634	283	86.4	11.2
NA	72.7	16.5	71.3	21.2	73	134	75.8	18.5	73	134	36.0	22.6
OA	75.2	10.2	78.6	10.3	669	74	48.7	1.5	677	126	65.8	13.7
GM	72.7	11.8	<b>75.8</b>	12.3	656	77	52.6	4.4	647	132	53.3	17.8
CMC												
PA	73.6	7.3	78.6	1.9	718	10	68.4	5.3	736	64	80.6	4.7
NA	57.3	4.4	56.4	5.1	608	37	60.9	5.6	557	78	52.6	4.3
OA	66.6	4.1	69.3	3.2	672	12	65.5	2.1	662	07	68.7	1.4
GM	64.8	3.2	<b>66.5</b>	3.2	66	18	64.4	2.0	637	20	65.0	0.9
Ionosphere												
PA	93.1	5.5	95.7	4.1	91	34	83.5	12.5	91	75	95.3	5.8
NA	91.9	4.8	92.8	6.0	886	68	91.3	9.6	917	35	85.2	7.8
OA	92.5	2.4	94.5	3.8	90	26	85.7	6.1	910	52	91.7	3.7
GM	92.94	1.6	<b>94.2</b>	3.9	897	34	86.8	4.8	91	41	90.0	4.2
Diabetes												
PA	74	3.9	76.6	4.2	625	44	57.2	16.5	76	25	82.3	8.7
NA	67	8.3	71	11.9	775	71	75.9	10.8	651	113	57.0	5.7
OA	71.5	4.6	74.4	6.5	671	23	64.5	6.5	725	54	73.8	6.2
GM	70.3	5.4	<b>73.6</b>	7.9	695	18	64.7	5.1	701	69	68.2	4.7

Note: "DT" means we do nothing for imbalanced data and just implement decision tree algorithm.

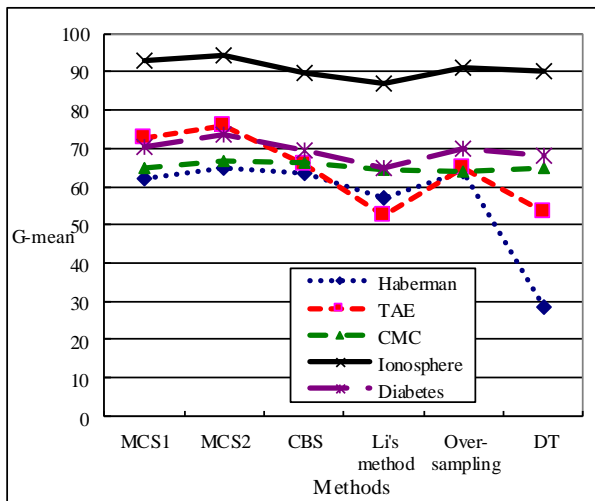


Fig. 2. Comparisons between the proposed MCS method and traditional methods by considering G-mean

D. Results of Three Imbalanced Sentiment Classification Cases

This section will evaluate the effectiveness of the proposed MCS1 and MCS2 in textual data. Table 4 summarizes the experimental results of implementing our methods in three real cases of sentiment classification. From the results of DT, we can find there are serious class imbalance problems in all three cases. It means that in these cases PA is very high but NA is unacceptably low. In other words, if we do nothing for these three imbalanced sentiment classification data, a classifier cannot identify any negative comments (minority class).

TABLE IV  
THE RESULTS OF THREE SENTIMENT CLASSIFICATION CASES

Index	Proposed Methods				Traditional Methods							
	MCS1 (%)		MCS2 (%)		CBS (%)		Li's method (%)		Over-sampling (%)		DT (%)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
MP3												
PA	82.1	11.9	78.2	15.2	34.1	8.8	40.9	28.2	87.4	5.6	97.9	3.1
NA	30.6	5.5	52.7	18.9	91.6	5.5	66.6	32.7	44.4	15.7	2.7	5.5
OA	75.4	10.6	74.8	10.8	41.5	8.4	44.1	20.9	81.8	3.0	85.7	2.0
GM	49.9	6.7	<b>62.3</b>	4.9	55.7	8.7	45.6	3.9	61.1	10.5	8.0	16.1
EC												
PA	72.0	17.3	81.3	18.9	74.1	33.1	66.3	36.1	80.1	15.2	96.7	3.8
NA	54.9	11.7	49.0	12.6	50.1	27.7	55.1	26.8	46.5	18.4	27.2	16.1
OA	68.1	11.9	71.2	17.3	66.3	17.5	58.8	17.8	71.3	8.5	73.1	12.1
GM	61.7	3.6	<b>63.0</b>	14.6	54.9	5.2	54.3	4.7	59.0	7.7	49.7	13.9
Movie												
PA	74.5	2.7	78.9	2.3	59.4	4.1	51.3	5.1	73.9	4.1	78.6	3.5
NA	45.1	4.2	39.4	6.7	53.1	4.6	63.7	6.7	41.3	3.9	36.8	2.7
OA	65.4	3.5	66.8	1.7	57.4	1.8	55.3	2.2	63.8	3.2	65.6	2.2
GM	<b>57.9</b>	3.6	55.6	4.2	56.1	1.5	56.9	1.9	55.2	3.6	53.8	2.6

In MP3 case, MCS2 has a better G-mean (62.39%) than MCS1 (49.9%), CBS (55.74%), Li's method (45.68%), over-sampling (61.15%) and DT (8.05%). Besides, in top three good methods, MCS2 is more stable (SD: 4.99%) than over-sampling (SD: 10.59%) and CBS (8.71%). In EC case, top two methods which has best performance are the proposed MCS1 (61.70%) and MCS2 (63.06%). Both of them

outperform CBS (54.9%), Li's method (54.3%), over-sampling (59.0%) and DT (49.7%). But, in this case, MCS1 (SD: 3.61%) is more stable than MCS2 (SD: 14.68%). Besides, compared with traditional re-sampling techniques including CBS, Li's method, and over-sampling, MCS1 and MCS2 do lose too much overall accuracy.

Figure 3 provides an overview of comparisons between the proposed MCS methods and traditional re-sampling methods in movie review case, MCS1 has better performance than MCS2, CBS, Li's method, and DT. Particularly, G-mean of MCS1 is 57.9%, of MCS2 is 55.62%, of CBS, of Li's method is 56.9%, over-sampling is 55.2%, and of DT 53.8%, respectively. Similarly, SD of MCS1 is 3.68%, of MCS2 is 4.62% of CBS is 1.51%, of Li's method is 1.95%, of over-sampling is 3.61%, and of is 2.69%, respectively.

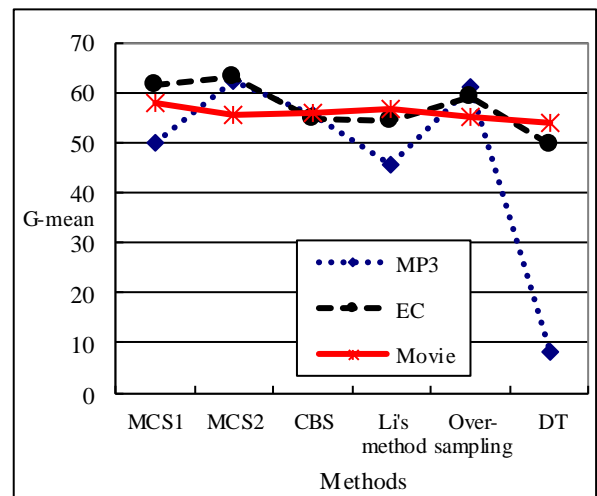


Fig. 3. Comparisons between the proposed MCS method and traditional methods in three sentiment classification cases

In order to testify the difference between proposed MCS1, MCS2 and traditional CBS, over-sampling, Li's method, we implement statistical hypotheses under 95% confidence level. Table 5 summarizes the results. From this table, we can make some concluding remarks: (1) MCS2 is significantly better than Li's method; (2) MCS1 is significantly better than Li's method; (3) the proposed MCS1 and MCS2 both significantly outperform DT which is without implementing any technique.

From these results, it can be concluded that MCS2 is slightly better than MCS1 since the difference is not significant. And both the proposed MCS1 and MCS2 are significantly superior to traditional methods in most cases. The evidences of implementing statistical hypotheses can support my concluding remarks.

TABLE V RESULTS OF HYPOTHESES TESTS FOR COMPREHENSIVE EVALUATIONS

No.	Hypothesis	G-mean	
		P-value	Conclusion
H1	$H_0 : \mu_{MCS2} \leq \mu_{MCS1}$ $H_1 : \mu_{MCS2} > \mu_{MCS1}$	0.266	Accept $H_0$
H2	$H_0 : \mu_{MCS2} = \mu_{MCS1}$ $H_1 : \mu_{MCS2} \neq \mu_{MCS1}$	0.533	Accept $H_0$
H3	$H_0 : \mu_{MCS2} \leq \mu_{Li's\ method}$ $H_1 : \mu_{MCS2} > \mu_{Li's\ method}$	<b>0.009</b>	<b>Reject <math>H_0</math></b>
H4	$H_0 : \mu_{MCS2} \leq \mu_{CBS}$ $H_1 : \mu_{MCS2} > \mu_{CBS}$	0.138	Accept $H_0$
H5	$H_0 : \mu_{MCS2} \leq \mu_{over-sampling}$ $H_1 : \mu_{MCS2} > \mu_{over-sampling}$	0.148	Accept $H_0$
H6	$H_0 : \mu_{MCS2} \leq \mu_{DT}$ $H_1 : \mu_{MCS2} > \mu_{DT}$	<b>0.019</b>	<b>Reject <math>H_0</math></b>
H7	$H_0 : \mu_{MCS1} \leq \mu_{Li's\ method}$ $H_1 : \mu_{MCS1} > \mu_{Li's\ method}$	<b>0.035</b>	<b>Reject <math>H_0</math></b>
H8	$H_0 : \mu_{MCS1} \leq \mu_{CBS}$ $H_1 : \mu_{MCS1} > \mu_{CBS}$	0.330	Accept $H_0$
H9	$H_0 : \mu_{MCS1} \leq \mu_{over-sampling}$ $H_1 : \mu_{MCS1} > \mu_{over-sampling}$	0.326	Accept $H_0$
H10	$H_0 : \mu_{MCS1} \leq \mu_{DT}$ $H_1 : \mu_{MCS1} > \mu_{DT}$	<b>0.041</b>	<b>Reject <math>H_0</math></b>

E. Effects of Small Sample Size Problem

This section will introduce the effects of the small sample size (SSS) problem which means the number of attributes (x) is larger than data size (n) on our proposed methods. Since the SSS problem is very common in textual data, we used three sentiment classification data to do experiments. In order to create an SSS situation, we reduced data size until the number of attributes (x) is larger than data size (n). In this study, we experimented two SSS situations in which x: n has been set up as 2:1 and 4:1, respectively. Table 6 provides the setting of the SSS situation in our experiments. Figures 4~6 list the results of different SSS situations on three textual data. It might not be easy to see the performance change when varying SSS situations. Therefore, we merely considered G-mean and drew classifiers' performance changes as in Figures 4~6.

TABLE VI THE SETTINGS OF SMALL SAMPLE SIZE SITUATIONS

Data \ Ratio	the amount of attributes : data size		
	Original (x : n)	Small sample size	
		2 : 1	4 : 1
EC	300 : 349	300 : 150	300 : 75
MP3	66 : 200	66 : 33	66 : 16
Movie	325 : 1000	325 : 162	325 : 81

Note: "x" means the amount of attributes and "n" is data size.

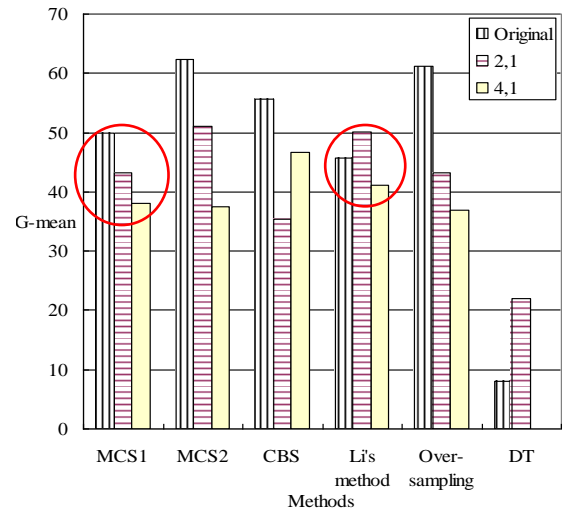


Fig. 4. The results of considering small sample size in MP3 case

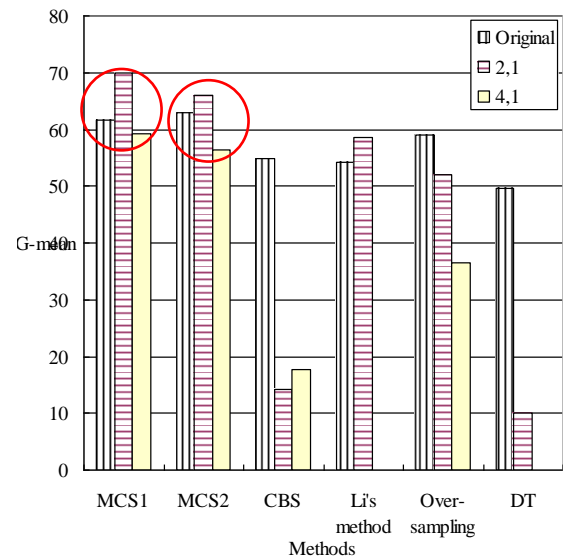


Fig. 5. The results of considering small sample size in EC case

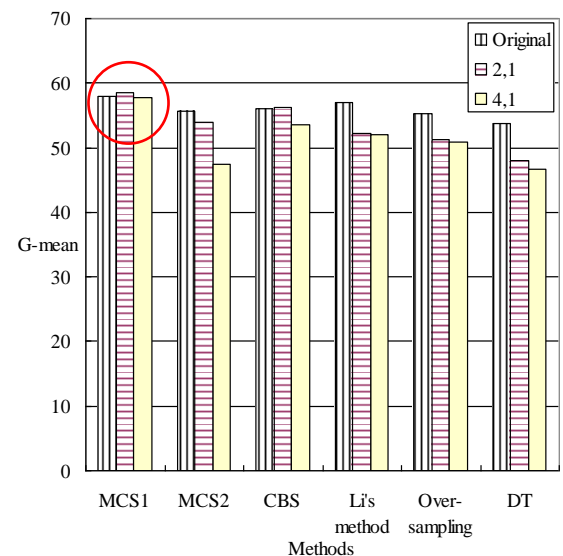


Fig. 6. The results of considering small sample size in Movie review case

Figure 4 shows the effects of the SSS problem in the MP3 cases. From this figure, we can find the performances of all



classifiers decrease except in Li's method. When the SSS situation changes from the original set to 4:1, Li's method is more stable than other methods. In addition, when only comparing MCS1 and MCS2, unlike MCS2, MCS1's performance drops slightly.

Figure 5 provides the effects of SSS problems in EC cases. From this figure, we can find that the performances of DT and Li's method decrease to zero when the SSS situation is set to 4:1. The results of over-sampling and traditional CBS also show that they suffer from SSS problems. In contrast, in our MCS1 and MCS2, their performances do not drop very dramatically. In contrast, they keep stable performances when varying SSS situations. Considering movie review results in Figure 6, MCS1 cannot be influenced when SSS situations change from original set to 4:1. However, the performances of other methods including MCS2 have decreased. To sum up, MCS1 outperforms MCS2 and other traditional re-sampling methods when thinking of SSS problems in textual sentiment data.

#### V. DISCUSSION AND CONCLUSIONS

The class imbalance problems are very common issues in text classification, so they need to be considered seriously when applying machine learning methods for sentiment classification. In order to enhance re-sampling methods, we proposed MCS1 and MCS2 methods which are expected to provide a possible solution for the class imbalance problems. The research data were five UCI data sets and three imbalanced sentiment classification data. After successful experiments, it can be concluded that in most cases, MCS2 outperforms MCS1 and other traditional techniques, such as CBS, Li's method, over-sampling, and DT. In addition, although some results (3 in 5 UCI data and 2 in 3 sentiment classification cases) indicate MCS1 is slightly better than other re-sampling methods, we have no significant evidence to claim that. This part is needed to be validated by doing additional experiments.

Secondly, SSS problems are very common and has become a critical issue in textual data. Without implementing dimensionality reduction approaches, MCS1 has a better and more stable performance than MCS2 and other traditional re-sampling techniques. The reason is that MCS2 uses an over-sampling method, which duplicates minority examples until the imbalanced situation are improved in order to select representative minority class examples. This approach might dramatically solve the SSS problems and better classification performance.

Integrating feature selection (feature extraction techniques) into the proposed method might be a potential direction for future works. Moreover, although this study focuses on imbalanced sentiment classification, our proposed method is also suitable for classifying sentiment in any text-based communication tools, such as text content-based social media and some websites combined with social commerce. However, if we need to identify customers' sentiment in a call center, voice data should be handled. It is also a good direction for future experiments.

#### REFERENCES

- [1] Abbasi, A., H. Chen, S. Thoms, T. Fu, Affect analysis of web forums and blogs using correlation ensembles, *IEEE Transactions on Knowledge and Data Engineering* 20(9) (2008) 1168-1180.
- [2] Alam, M., Abid, F., Guangpei, C., Yunrong, L. V., Social media sentiment analysis through parallel dilated convolutional neural network for smart city applications, *Computer Communications*, 154 (2020) 129-137.
- [3] Altincay, H. and C. Ergun, Clustering based under-sampling for improving speaker verification decisions using AdaBoost, *Lecture Notes in Computer Science* 3138 (2004) 698-706.
- [4] Andrzej Z. Grzybowski, "Classification Problems with Unequal Error Costs - Performance of Selected Global Optimization Algorithms," *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2019, 22-24 October, 2019, San Francisco, USA*, pp293-297
- [5] Argamon, S., M. Koppel, and G. Avneri, Routing documents according to style, *Proceeding of First international workshop on innovative information systems* (1998).
- [6] Asif, M., Ishtiaq, A., Ahmad, H., Aljuaid, H., Shah, J., Sentiment analysis of extremism in social media from textual information, *Telematics and Informatics*, 48 (2020) Article 101345.
- [7] Batista, G. E. A. P. A., R. C. Prati, and M. C. Monard, A study of the behaviour of several methods for balancing machine learning training data, *SIGKDD Explorations* 6(1) (2004) pp. 20-29.
- [8] Bria, A., Marrocco, C., Tortorella, F., Addressing class imbalance in deep learning for small lesion detection on medical images, *Computers in Biology and Medicine*, 120 (2020) Article 103735.
- [9] Chang, J.-R., Chen, M.-Y., Chen, L.-S., Chien, W.-T., Recognizing Important Factors of Influencing Trust in O2O Models: An Example of OpenTable, *Soft Computing*, 24 (2020) 7907-7923.
- [10] Chang, J.-R., Chen, M.-Y., Chen, L.-S., Tseng, S.-C., Why Customers Don't Revisit in Tourism and Hospitality Industry? *IEEE Access*, 7(7) (2019) 146588-146606.
- [11] Chaovalit, P. and L. Zhou, Movie review mining: A comparison between supervised and unsupervised classification approaches, *Proceedings of IEEE international conference on system sciences* (2005) 1-9.
- [12] Chau, M. & Xu, J. (2007). Mining communities and their relationships in blogs: a study of online hate groups. *International Journal of Human - Computer Studies*, 65(1), 57-70.
- [13] Chen, E., Lin, Y., Xiong, H., Luo, Q., Ma, H., Exploiting probabilistic topic models to improve text categorization under class imbalance, *Information Processing and Management*, 47(2) (2011) 202-214.
- [14] Chen, L. S., C. C. Hsu and M. C. Chen, Customer segmentation and classification from blogs by using data mining: An example of VOIP phone, *Cybernetics & Systems* 40(7) (2009) 608-632.
- [15] Chen, L. S., C. C. Hsu, Y. S. Chang, MDS: A novel method for class imbalance learning, *The 3rd International Conference on Ubiquitous Information Management and Communication*, January 15-16, 2009a, SKKU, Suwon, Korea.
- [16] Chen, L. S., Y. S. Chang, L. W. Lin, M. C. Chen, , A SOM based approach for learning from imbalanced data sets, *The 19th Intelligent System Symposium (FAN 2009)*, Sep. 17-18, 2009b, Aizu-Wakamatsu, Japan.
- [17] Chen, M. C., L. S. Chen, C. C. Hsu, and W. R. Zeng, An information granulation based data mining approach for classifying imbalanced data, *Information Sciences* 178(16) (2008) 3214-3227.
- [18] Cohen, E., and B. Krishnamurthy, A short walk in the Blogistan, *Computer Networks* 50(5) (2006) 615-630.
- [19] Cohen, G., M. Hilario, H. Sax, S. Hugonnet, and A. Geissbuhler, Learning from imbalanced data in surveillance of nosocomial infection, *Artificial Intelligence in Medicine* 37 (2006) 7-18.
- [20] Dave, K., S. Lawrence, and D. M. Pennock, Mining the peanut gallery: opinion extraction and semantic classification of product reviews, *The 12th WWW* (2003) 519-528.
- [21] Denecke, K. and W. Nejdl, How valuable is medical social media data? Content analysis of the medical web, *Information Sciences* 179 (2009) 1870-1880.
- [22] El-Alfy, E. M., Al-Azani, S., Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short Arabic text, *Procedia Computer Science*, 109 (2017) 359-366.
- [23] Estabrooks, A., T. Jo, and N. Japkowicz, A multiple resampling method for learning from imbalanced data sets, *Computational Intelligence* 20(1) (2004) 18-36.

- [24] Fernández, A., M. Jesus, F. Herrera, On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data-sets, *Information Sciences* 180 (2010) 1268-1291.
- [25] Fu, G.-H., Wu, Y.-J., Zong, M.-J., Yi, L.-Z., Feature selection and classification by minimizing overlap degree for class-imbalanced data in metabolomics, *Chemometrics and Intelligent Laboratory Systems*, 196 (2020) Article 103906.
- [26] García, V., Sánchez, J. S., Marqués, A. I., Florencia, R., Rivera, G., Understanding the apparent superiority of over-sampling through an analysis of local information for class-imbalanced data, *Expert Systems with Applications*, In press, (2019) Article 113026, <https://doi.org/10.1016/j.eswa.2019.113026>
- [27] Gokalp, O., Tasci, E., Ugur, A., A novel wrapper feature selection algorithm based on iterated greedy metaheuristic for sentiment classification, *Expert Systems with Applications*, 146 (2020) Article 113176.
- [28] Guo, H. and H. L. Viktor, Learning from imbalanced data sets with boosting and data generation: the DataBoost- IM approach, *SIGKDD Explorations* 6(1) (2004) 30-39.
- [29] Hamidzadeh, J., Kashefi, N., Moradi, M., Combined weighted multi-objective optimizer for instance reduction in two-class imbalanced data problem, *Engineering Applications of Artificial Intelligence*, 90 (2020) Article 103500.
- [30] He, H. and E. A. Garcia, Learning from Imbalanced Data, *IEEE Transactions on Knowledge and Data Engineering* 21(9) (2009) 1263-1284.
- [31] Hendry, and Chen, R.-C., Using Deep Learning to Predict User Rating on Imbalance Classification Data, *IAENG International Journal of Computer Science*, 46(1) (2019) 109-117.
- [32] Henriquez, C., Briceno, F., and Salcedo, D., Unsupervised model for aspect-based sentiment analysis in Spanish, *IAENG International Journal of Computer Science* 46:3 (2019) 430-438.
- [33] Huang, K., H. Yang, I. King and M. Lyu, Learning classifiers from imbalanced data based on biased mini-max probability machine, *Proceedings of the 04' IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2004) 558-563.
- [34] Hulse, J. V., T. M. Khoshgoftaar, A. Napolitano, An empirical comparison of repetitive undersampling techniques, *Proceeding of IEEE International Conference on Information Reuse & Integration* (2009) 29-34.
- [35] Ibrahim, A. A., Abdulaziz, R. O., Analysis of Titanic Disaster using Machine Learning Algorithms, *Engineering Letters*, 28(4) (2020) 1161-1167.
- [36] Japkowicz, N. and S. Stephen, The class imbalance problem: a systematic study, *Intelligent Data Analysis* 6(5) (2002) 429-449.
- [37] Jo, T. and N. Japkowicz, Class imbalances versus small disjuncts, *SIGKDD Explorations* 6(1) (2004) 40-49.
- [38] Joachims, T., Making large-scale SVM learning practical, In B. Scholkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in kernel methods: Support vector learning*. The MIT Press, (1998), 37
- [39] Kamalov, F., Kernel density estimation based sampling for imbalanced class distribution, *Information Sciences*, 512 (2020) 1192-1201.
- [40] Kessler, B., G. Nunberg, and H. SchAutze, Automatic detection of text genre, *Proceedings of the 35th ACL/8th EACL* (1997) 32-38.
- [41] Li, P., P. L. Qiao, and Y. C. Liu, A hybrid re-sampling method for SVM learning from imbalanced data sets, *Fifth International Conference on Fuzzy Systems and Knowledge Discovery* (2008) 65-69.
- [42] Li, Y., Guo, H., Zhang, Q., Gu, M., Yang, J., Imbalanced text sentiment classification using universal and domain-specific knowledge, *Knowledge-Based Systems*, 160 (2018) 1-15.
- [43] Liu, B., Tsoumakas, G., Dealing with class imbalance in classifier chains via random undersampling, *Knowledge-Based Systems*, 192 (2020) Article 105292.
- [44] Liu, J., Q. Hu, D. Yu, A weighted rough set based method developed for class imbalance learning, *Information Sciences* 178(4) (2008) 1235-1256.
- [45] Liu, Y., Loh, H. T., Sun, A., Imbalanced text classification: A term weighting approach, *Expert Systems with Applications*, 36 (2009) 690-701.
- [46] Manevitz, L. M. and M. Yousef, One class SVMs for document classification, *Journal of Machine Learning Research* 2 (2002) 139-154.
- [47] Min, F. and Q. Liu, A hierarchical model for test-cost-sensitive decision systems, *Information Sciences* 179 (2009) 2442-2452.
- [48] Moraes, R., Valiati, J. F., Neto, W. P. G., Document-level sentiment classification: An empirical comparison between SVM and ANN, *Expert Systems with Applications*, 40 (2) (2013) 621-633.
- [49] Nakayama, M., Wan, Y., The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews, *Information & Management*, 56(2) (2019) 271-279.
- [50] Ofek, N., Rokach, L., Stern, R., Shabtai, A., Fast-CBUS: A fast clustering-based undersampling method for addressing the class imbalance problem, *Neurocomputing*, 243 (2017) 88-102.
- [51] Ogura, H., Amano, H., Kondo, M., Comparison of metrics for feature selection in imbalanced text classification, *Expert Systems with Applications*, 38(5) (2011) 4978-4989.
- [52] Orriols-Puig, A., E. Bernadó-Mansilla, D. E. Goldberg, K. Sastry, and P. L. Lanzi, Facetwise analysis of XCS for problems with class imbalances, *IEEE Transactions on Evolutionary Computation* 13(5) (2009) 1093-1119.
- [53] Padurariu, C., Breaban, M. E., Dealing with Data Imbalance in Text Classification, *Procedia Computer Science*, 159 (2019) 736-745.
- [54] Pang, B., L. Lee, and S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, *EMNLP* (2002) 79-86.
- [55] Qian, W., Li, S., A novel class imbalance-robust network for bearing fault diagnosis utilizing raw vibration signals, *Measurement*, 156 (2020) Article 107567.
- [56] Qiu, J., Liu, C., Li, Y., Lin, Z., Leveraging sentiment analysis at the aspects level to predict ratings of reviews, *Information Sciences*, 451-452 (2018) 295-309.
- [57] Raghuvanshi, B.S., Shukla, S., SMOTE based class-specific extreme learning machine for imbalanced learning, *Knowledge-Based Systems*, 187(2020) Article 104814.
- [58] Richhariya, B., Tanveer, M., A reduced universum twin support vector machine for class imbalance learning, *Pattern Recognition*, 102 (2020) Article 107150.
- [59] Ruz, G. A., Henríquez, P. A., Mascareño, A., Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers, *Future Generation Computer Systems*, 106 (2020) 92-104.
- [60] Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Spyropoulos, C., Atamatopoulos, P., Stacking classifiers for anti-spam filtering of E-mail, in: L. Lee, D. Harman (Eds.), *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, Carnegie Mellon University, 2001, 44-50.
- [61] Singh, T., L. Veron-Jackson, J. Cullinane, Blogging: A new play in your marketing game plan, *Business Horizons* 51(4) (2008) 281-292.38
- [62] Spertus, E., Smokey: Automatic recognition of hostile messages, *Proceeding of IAAI* (1997).
- [63] Stamatatos, E., Author identification: Using text sampling to handle the class imbalance problem, *Information Processing and Management*, 44 (2008) 790-799.
- [64] Su, C. T., L. S. Chen, and T. L. Chiang, A neural network based information granulation approach to shorten the cellular phone test process, *Computers In Industry* 57(5) (2006) 412-423.
- [65] Su, C. T., L. S. Chen, and Y. Yih, Knowledge acquisition through information granulation for imbalanced data, *Expert System with Applications* 31(3) (2006) 531-541.
- [66] Su, C. T., Y.-H. Hsiao, An evaluation of the robustness of MTS for imbalanced data, *IEEE Transaction on Knowledge and Data Engineering* 19(10) (2007) 1321-1332.
- [67] Sun, A., Lim, E.-P., Liu, Y., On strategies for imbalanced text classification using SVM: A comparative study, *Decision Support Systems* 48 (2009) 191-201.
- [68] Suresh, S., N. Sundararajan, P. Saratchandran, Risk-sensitive loss functions for sparse multi-category classification problems, *Information Sciences* 178(12) (2008) 2621-2638.
- [69] Tian, F., Wu, F., Chao, K.-M., Zheng, Q., Shah, N., Lan, T., Yue, J., A topic sentence-based instance transfer method for imbalanced sentiment classification of Chinese product reviews, *Electronic Commerce Research and Applications*, 16, (2016) 66-76.
- [70] Vashishtha, S., Susan, S., Fuzzy rule based unsupervised sentiment analysis from social media posts, *Expert Systems with Applications*, 138 (2019) Article 112834.
- [71] Wang, S., Li, D., Zhao, L., Zhang, J., Sample cutting method for imbalanced text sentiment classification based on BRC, *Knowledge-Based Systems*, 37 (2013) 451-461.
- [72] Weiss, G. and F. Provost, Learning when training data are costly: The effect of class distribution on tree induction, *Journal of Artificial Intelligence Research* 19 (2003) 315-354.
- [73] Weiss, G. and F. Provost, The effect of class distribution on classifier learning, *Technical Report, MLTR43*, Department of Computer Science, Rutgers University, (2001).

- [74] Wiebe, J. and E. Riloff, Creating subjective and objective sentence classifiers from unannotated texts, *Proceeding of Sixth international conference on intelligent text processing and computational linguistics* (2005).
- [75] Williams, D. P., V. Myers, and M. S. Silvious, Mine classification with imbalanced data, *IEEE Geoscience and Remote Sensing Letters*, 6(3) (2009) 528-532.
- [76] Wilson, T., J. Wiebe, and P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, *Proceedings of human language technologies conference/conference on empirical methods in natural language processing (HLT/ EMNLP 2005)*, Vancouver, Canada.
- [77] Wong, M. L., Seng, K., Wong, P. K., Cost-sensitive ensemble of stacked denoising autoencoders for class imbalance problems in business domain, *Expert Systems with Applications*, 141 (2020) Article 112918.
- [78] Wu, C. H., Z. J. Chuang, and Y. C. Lin, Emotion recognition from text using semantic labels and separable mixture models, *ACM Transactions on Asian Language Information Processing* 5(2) (2006) 165-183.
- [79] Wu, G. and E. Y. Chang, KBA: kernel boundary alignment considering imbalanced data distribution, *IEEE Transactions on Knowledge and Data Engineering* 17(6) (2005) 786-795.
- [80] Xi, P.-P., Zhao, Y.-P., Wang, P.-X., Li, Z.-Q., Pan, Y.-T., Song, F.-Q., Least squares support vector machine for class imbalance learning and their applications to fault detection of aircraft engine, *Aerospace Science and Technology*, 84 (2019) 56-74.
- [81] Xu, F., Pan, Z., Xia, R., E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework, *Information Processing & Management*, In press, (2020) Article 102221.
- [82] Yan, L., D. Xie, Z. Du, A new method of support vector machine for class imbalance problem, *2009 International Joint Conference on Computational Sciences and Optimization* (2009) 904-907.
- [83] Zanin, L., Combining multiple probability predictions in the presence of class imbalance to discriminate between potential bad and good borrowers in the peer-to-peer lending market, *Journal of Behavioral and Experimental Finance*, 25 (2020) Article 100272.
- [84] Zhang, D., Xu, H., Su, Z., Xu, Y., Chinese comments sentiment classification based on word2vec and SVMperf, *Expert Systems with Applications*, 42(4) (2015) 1857-1863.
- [85] Zhang, Y., Miao, D., Wang, J., Zhang, Z., A cost-sensitive three-way combination technique for ensemble learning in sentiment classification, *International Journal of Approximate Reasoning*, 105 (2019) 85-97.
- [86] Zheng, Z., Wu, X., Srihari, R., Feature selection for text categorization on imbalanced data, *ACM SIGKDD Explorations Newsletter-Special issue on learning from imbalanced datasets*, 6(1) (2004) 80-89.

**Dr. Jing-Rong Chang** is an Associate Professor with Department of Information Management at Chaoyang University of Technology, Taiwan. He received his Ph.D. from College of Management, National Yunlin University of Science and Technology, Taiwan in 2006, and his MS degree in department of information management from National Yunlin University of Science and Technology in 2002. His teaching and research interests include data mining, soft computing, decision analysis and intelligent systems.

**Dr. Long-Sheng Chen** is a Professor with Department of Information Management, and Assistant Vice President for Academic Affairs, Chaoyang University of Technology, Taiwan. He received his Ph.D. in Department of Industrial Engineering and Management, National Chiao Tung University, Taiwan in 2006, and his BS and MS degrees both in industrial management from National Cheng Kung University, Tainan, Taiwan in 1998 and 2000, respectively. His current research interests include data mining, text mining, social media, quality management, and customer relationship management, context-awareness, and machine learning, with more than 100 publications in these areas.

**Mr. Li-Wei Lin** is a graduate student of Department of Information Management, Chaoyang University of Technology, Taiwan. He received his MS degree in Department of Information Management, Chaoyang University of Technology, Taiwan in 2010. His current research interests include data mining and classification.