Identifying Key Factors and Developing a New Method for Classifying Imbalanced Sentiment Data

Long-Sheng Chen* and Kun-Cheng Sun

***Abstract*—Bloggers' opinions related to commercial products/services might have a significant influence on consumers' purchasing decisions. Some negative comments could reduce consumers' purchase intentions and bring a great damage to enterprises. But, the comments in blogs are often unstructured, subjective, and hard to comprehend in short time. In some cases, the negative comments are usually fewer than the positive opinions. These fewer negative comments spread very fast and are much harmful. According to a consumer reviews and research online report, 62% of online customers will change their mind about buying a product or service after reading 1~3 negative reviews. But, when dealing with such imbalanced sentiment data, researchers didn't consider the class imbalance problem. A classifier induced from an imbalanced data set has high classification accuracy for the majority class, but an unacceptable error rate for the minority class. Therefore, to identify consumers' negative sentiments effectively from a large number of online comments had become one of serious issues. So, this study aims to identify the key factors of imbalanced sentiment classification by using Taguchi method. Then, according to the discovered key factors, we'll propose a new feature selection method to improve the performance of imbalanced sentiment classification. Moreover, support vector machines (SVM) have been employed to construct classifiers for identifying bloggers' negative sentiments. Finally, one case study from real world blogs will be provided to illustrate the effectiveness of our proposed approach.**

***Index Terms*—Class imbalance problems, Sentiment classification, Feature selection, Text classification, Taguchi methods.**

I. Introduction

In recent decades, electronic communication has changed to be more convenient and ubiquitous. Once E-mail has been a staple, but as the Internet grew there has been FTP, BBS, web pages, secured servers and now the world of wikis, blogs (i.e. weblogs) [10], YouTube, Twitter, and Facebook. They have become crucial media for expressing personal opinions, sharing information, and communicating to each other [3,12]. Among them, blogs are one of the fastest growing sections of the Internet and are emerging as an important communication mechanism that is used by an increasing number of people [3, 4]. Blogs also have been regarded as the 4th Internet application which can cause radical changes in the world, after E-mail, Instant Message, and BBS. Rosenbloom [9] indicates that blogs are becoming a new form of mainstream personal communication for millions of people to publish and exchange knowledge/information, and to establish networks or build relationships in the world of all blogs, so-called "blogsphere". According to a survey of Nielsen in 2010, internet users spent 22 percent of all time online on social networks and blogs [8]. Therefore, bloggers' comments have great power for their readers. Some personal opinions related to commercial products/services might have a significant influence on consumers' purchasing decisions [5, 11]. Channel Advisor cooperation made a customer shopping habit survey that found 92% online users will read related comments before buying a product [1]. So, these comments can provide product information and recommendations from the customer perspectives [6]. Besides, the comments in blogs are often unstructured, subjective, and hard to comprehend in short time. Consequently, to identify consumers' sentiments effectively from a large number of online comments had become one of serious issues [2].

However, some comments are usually negative and they could reduce consumers' purchase intentions. In some cases, the negative comments are usually fewer than the positive opinions. But, these fewer negative comments will spread very fast and bring a great damage to enterprises. According to a consumer reviews and research online report of Lightspeed company [7], 62% of online customers will change their mind about buying a product or service after reading 1~3 negative reviews. But, when dealing with such imbalanced sentiment data, researchers didn't consider the class imbalance problem (lots of bloggers' comments are positive and far fewer comments are negative). A classifier induced from an imbalanced data set has high classification accuracy for the majority class, but an unacceptable error rate for the minority class.

In related works of sentiment classification, feature selection is one of effective method to improve the sentiment classification. However, Zheng et al. [26] found traditional feature selection methods tend to select features from majority examples, and they presented an index named Signed IG which combined sign index and IG to deal with class imbalance problems in text categorization. Ogura et al.

Manuscript received December 30, 2011. This work was supported in part by the National Science Council of Taiwan, R.O.C.

*L.-S. Chen is an associate professor with Department of Information Management, Chaoyang University of Technology, Taichung 41349, Taiwan (phone: 886-4-23323000ext7752; fax: 886-4-23304902; e-mail: lschen@cyut.edu.tw).

K.-C. Sun is a graduate student of Department of Information Management, Chaoyang University of Technology, Taichung 41349, Taiwan (e-mail:s9914641@cyut.edu.tw).

[27] used signed IG and signed CHI for imbalanced text data. But, the signed feature selection methods still cannot remarkably improve the performance of classifying imbalanced sentiment data.

Therefore, this study aims to identify the key factors of imbalanced sentiment classification by using Taguchi method. Then, according to the discovered key factors, we'll propose a new feature selection method which equally selects features from both positive and negative sentiments to improve the performance of imbalanced sentiment classification. Moreover, support vector machines (SVM) have been employed to construct classifiers for identifying bloggers' negative sentiments. Finally, one case study from real world blogs will be provided to illustrate the effectiveness of our proposed approach. Compared with traditional methods, experimental results indicated that the proposed method can improve sentiment classification performance.

II. Related works

A. Sentiment Classification

Recently, sentiment classification that classifies bloggers' opinions into negative and positive groups has attracted lots of attention in web mining area [16]. Generally speaking, the objective of sentiment classification is to extract opinions from customers for certain products or services, and to identify reviews' sentiment [22]. Sentiment classification can detect bloggers' emotions to assist companies to carefully respond to customers' comments. In available works, two groups of popular approaches have been employed to solve this issue [18, 19, 23]. They are machine learning methods and information retrieval techniques.

Machine learning techniques attempt to build classifiers from sentiment labeled textual comments, and then identify the sentiment of new coming comments in blogs based on this constructed classifiers. Information retrieval approaches classify terms into two classes (positive or negative), and then count the overall positive and negative scores in the documents to determine the sentiment of comments [21]. According to lots of published literatures, machine learning methods have been considered as one of effective solutions for sentiment classification. For examples, Dave et al. [20] develop a method of classifying positive and negative reviews automatically and experiment several methods related to feature selections and scoring. Whitelaw et al. [17] presented a new method for sentiment classification based on extracting and analyzing appraisal groups with Support Vector Machines (SVM) classifier. Abbasi et al. [13] developed Entropy Weighted Genetic Algorithm (EWGA) feature selection method with Support Vector Machines (SVM) classifier for sentiment classification on movie reviews. Abbasi et al. [14] proposed SVRCE approach to analyze emotional states. O'Keefe and Koprinska [25] compared three feature selection methods under considering a number of selection thresholds and using six term weighting methods with both Naive Bayes and SVM classifiers. Although these studies indicated that using machine learning techniques could have a good performance for classifying textual sentiment data, but the high dimensionality of textual data will degrade the performance of classifiers and lead to long training time [24]. How to effectively and easily reduce the dimensionality of textual data and maintain the classification performance needs to be addressed.

B. Feature selection

Feature selection aims to extract relevant attributes to describe collected documents from a huge amount of candidate features and achieve a goal of dimension reduction in a short term. In general, feature selection technique has widely used to reduce dimensionality of textual data, to decrease computational time, and to remove irrelevant attributes and noise for improving classification performance [15]. Unlike feature selection algorithms for dealing with numerical data, such as Genetic Algorithm (GA) and Support Vector Machine Recursive Feature Extraction (SVM-RFE), which can result in good performance but they also need lots of computational cost. For text data, we need other feature selection methods to quickly select important terms and then construct term-document matrix based on them.

In related works, lots of methods have been presented for dimension reduction in text classification. In most of cases, researchers merely set a threshold of DF (document frequency) or TF (term frequency). If a term whose DF or TF is lower than this threshold, this term will be removed. Some researchers attempted to use POS tagging to select important attributes for classifying sentiment. However, until now, this kind approach cannot lead to a significant classification performance improvement [28, 29].

Other methods are to compute a score for each individual features and then pick out a predefined size of feature set according to the rank of scores, such as Chi-square statistic, mutual information, information gain and so on [22, 23, 25, 30, 31]. Zheng et al. [26] divided these feature selection methods into one-sided (eg. correlation coefficient and odds ratios) and two-sided (eg. information gain and Chi-square) groups. Among these methods, information gain (IG) is the most popular method and it has been proved effective in text classification. For examples, Tan and Zhang [23] compared 4 methods, document frequency (DF), mutual information (MI), Information gain (IG) and Chi-square (CHI), and their results indicated that IG outperformed other three methods when using support vector machines (SVM). Ye et al. [22] also employed IG as feature selection method and combined it into SVM, Naïve Bayes, and N-gram model for classifying sentiments of internet travellers' comments. Wang et al. [24] presented an improved Fisher's discriminant ratio for feature selection. Zheng et al. [26] presented a Sign index to deal with class imbalance problems in text categorization.

C. Taguchi method

Taguchi method has become one of well-recognized approaches to analyze the interaction effects when screening various controllable factors [35]. It can be used to determine the significant factors [36]. Unlike design of experiment (DOE), it only conducts the balanced (orthogonal) experimental combinations, which makes Taguchi method even more effective than a fractional factorial design [34]. By using the Taguchi method, we can remarkably reduce the cost

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

of experiments. Therefore, Taguchi method has been employed to find significant factors of imbalanced sentiment classification.

D. Support vector machines

SVM is a machine learning technique based risk minimization principle of statistical learning theory introduced by Vapnik [32], and it can deal with the problem of classification for multi-class or binary class. In the domain of sentiment classification, SVM aims to tackle the two-class problem by finding a hyperplane of maximal margin. Several studies [22, 23, 25] reported that SVM had a superior performance on sentiment classification. For this reason, we use SVM to be the classifier in our proposed feature selection method. Besides, the LIBSVM which was developed by Chang and Lin [33] has been employed to build SVM classifier.

III. Methodology

This section will introduce the implemental procedure of Taguchi method and our proposed method. In fact, there are two phases in this study. In phase 1, we attempt to identify key factors of imbalanced sentiment classification by using Taguchi method. Then, according to the discovered key factors, we'll propose a new feature selection method to improve the performance of imbalanced sentiment classification.

A. Taguchi Method

First, we introduce the implemental procedure of Taguchi method. Actually, there are 3 major steps which have been provided as follows.

Step 1: Planning the experiment

(1) Defining the control factors, noise factors and quality responses for the product or process.
(2) Determining the levels of each factor.
(3) Selecting an appropriate orthogonal array (OA) table. The selection of the most appropriate OA depends on the number of factors and interactions, and the number of levels of the factors.
(4) Transforming the data from the experiments into a proper S/N ratio.

Step 2: Implementing the experiment

Step 3: Analyzing and examining the results

(1) Using ANOVA analysis to determine the significant parameters.
(2) Using main effect plot analysis to determine the optimal levels of the control factors.
(3) Performing a factor contribution rate analysis.
(4) Confirming the experiment and planning future applications.

B. The proposed feature selection methodology

Before introducing our method, we need to address the question about how to determine positive and negative features. Let $d_{P,i}(1,2,...,m)$ and $d_{N,j}(1,2,...,n)$ represent the ith positive document and the jth negative document respectively. Random variables $d_{P,i}(t_k)$ and $d_{N,j}(t_k)$ are defined as equations (1) and (2).

$$d_{P,i}(t_k)=\begin{cases}1 & \text{if } t_k \text{ occurs in } \mathrm{d_{P,i}}\\ 0 & \text{otherwise}\end{cases} \quad (1)$$

$$d_{N,j}(t_k)=\begin{cases}1 & \text{if } t_k \text{ occurs in } \mathrm{d_{N,j}}\\ 0 & \text{otherwise}\end{cases} \quad (2)$$

Table 1 Two-way contingency table of a term t_k and a class C_j for binary classification

	Containing term t_k	Not containing term t_k
Belonging to class C_j	A	B
Not belonging to class C_j	C	D

To clearly illustrate how to separate features into positive and negative sets, we use notations in Table 1 to define related equations. In this table, notations A~B can be defined as equations (3)~(6). According to Zheng et al. [24], a feature's sign can be calculated as equation (7). If one feature's Sign value is >0 (<0), then this feature will be considered as "positive" ("negative").

$$A=\sum_{i=1}^{m} d_{P,i}(t_k) \quad (3)$$

$$B=m-\sum_{i=1}^{m} d_{P,i}(t_k) \quad (4)$$

$$C=\sum_{j=1}^{n} d_{N,j}(t_k) \quad (5)$$

$$D=n-\sum_{j=1}^{n} d_{N,j}(t_k) \quad (6)$$

$$Sign=AD-BC \quad (7)$$

Next, we introduce the implemented procedure of our proposed feature selection method which consists of following 6 steps. The flowchart is shown as Figure 1. The concise steps have been given as follows.

Step 1: Data collection and pre-process

We use unigram to represent collected documents. After removing some stop words, a set of candidate features can be constructed.

Step 2: Divide candidate features into positive and negative sets

According to equation (7), we calculate Sign value for every feature, and then assign those features whose Sign value is >0 (<0) to positive set F_+ (negative set F_-).

Step 3: Compute feature selection metrics

For F_+ and F_- sets, calculate each term's IG. Information gain (IG) has widely used as a term goodness criterion in the field of text classification [17]. Related literatures reported that IG can have good performances and have been viewed as one of the most effective feature selection methods. IG measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document. For a term t_k, its information gain can be defined as following equation (8).

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$IG(t_k) = H(C) - H(C|t_k)$$
$$= -\sum_{i=1}^{m} p(c_i)\log(p(c_i)) + p(t_k)\sum_{i=1}^{m} p(c_i|t_k)\log(p(c_i|t_k))$$
$$+ p(\bar{t}_k)\sum_{i=1}^{m} p(c_i|\bar{t}_k)\log(p(c_i|\bar{t}_k)) \quad (8)$$
$$= \sum_{i=1}^{m}\left(p(c_i,t_k)\log\left(\frac{p(c_i,t_k)}{p(c_i)p(t_k)}\right) + p(c_i,\bar{t}_k)\log\left(\frac{p(c_i,\bar{t}_k)}{p(c_i)p(\bar{t}_k)}\right)\right)$$

where $p(c_i)$ is the probability that category c_i occurs, $p(t_k)$ is the probability that term t_k occurs, $p(\bar{t}_k)$ denotes the probability that term t_k does not occur, $p(c_i,t_k)$ means the joint probability of c_i and t_k, and $p(c_i,\bar{t}_k)$ represents the joint probability of c_i and $\bar{t}_k$.

Next, one feature's IG is multiplied by it Sign value. Then, we can obtain a feature's Sign-IG score.

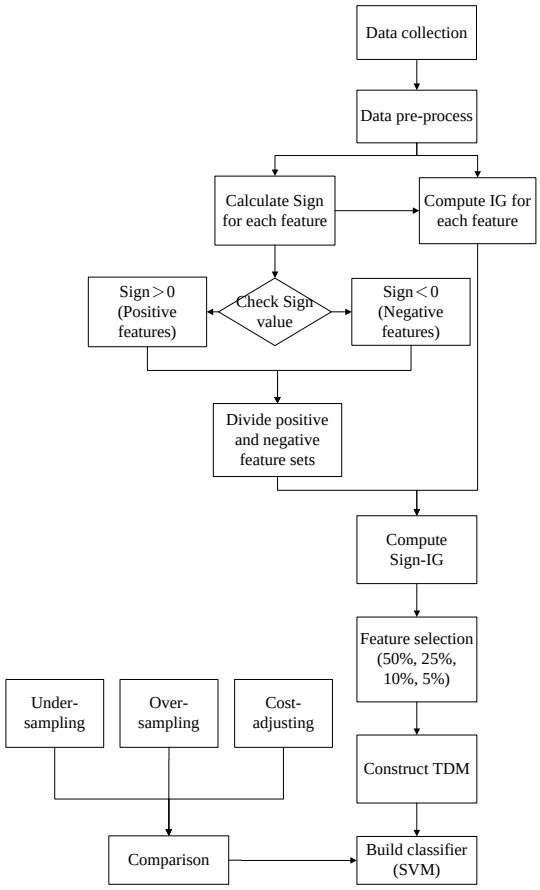

Figure 1 The flowchart of the proposed feature selection method

Step 4: Feature selection

We rank the computed Sign-IG scores in F_+ and F_-, respectively. Then, according to the rank of Sign-IG scores, we select important attributes equally from both F_+ and F_- based on the pre-determined dimension size. Then, join the selected features of F_+ and F_- to be the employed features of training data. In this work, we will reduce the dimension size to 50%, 25%, 10%, 5%, respectively, from original dimension size.

Step 5: Construct term-document matrix

TF-IDF (term frequency-inverse document frequency) will be employed to denote the term weights. Based on selected features in step4, the collected documents will be transformed to a term-document matrix (TDM). In addition, five-fold cross validation experiment has been utilized in this study.

Step 6: Train SVM classifier and validate experiment result

In this step, we use the training data to construct SVM classifier. Several conventional techniques for dealing with imbalanced data such as under-sampling, over-sampling, and cost-adjusting will be implemented. The one that has the best performance will be integrated into our proposed feature selection method.

IV. Experiment Result

A. Results of Taguchi method

In this study, we select 6 controllable factors. Table 2 summarizes their definitions and levels. In factor A, we want to know the text data type (sentiment data and non-sentiment data) will influence the classification performance or not. Factor B is imbalance ratio, which represent the skewed situation of class distribution, defined by "the amount of majority examples/ the amount of minority examples". Factor C is dimension size. Factor D is different term weights, TF and TF-IDF. Factor E is the common techniques to deal with imbalanced data, under-sampling and over-sampling. The last factor F is type of classifiers, DT (decision trees) and SVM.

Table 2 Defined factors and their levels

Factors		Level 1	Level 2
A	Data type	Sentiment data	Non-sentiment data
B	Imbalance Ratio	3	9
C	Dimension size	Original (100%)	Reduced (50%)
D	Term weights	TF	TF-IDF
E	Sampling method	Under-sampling	Over-sampling
F	Classifier	SVM	DT

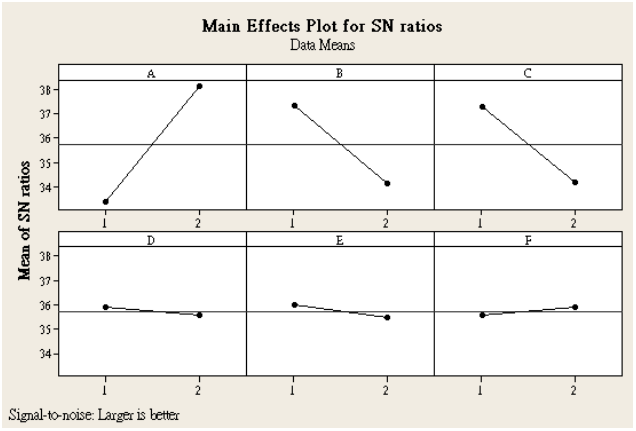

Figure 2 Main effect plot

Figure 2 provides the main effect of 6 factors. From this figure, we can find factors A, B, and C are significant for the imbalanced sentiment performance. Table 3 summarizes the detailed information of the analysis of variance. In this table, we combine non-significant factors D~F to the error item. From this table, we can find factor A (text data type) can mainly influence the performance. Imbalance ratio (factor B)

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

is the second place. The last significant factor is dimension size. But, since we focus on sentiment data and we cannot control the imbalance ratio of collected data, we try to develop feature selection method to reduce dimension size.

Table 3 ANOVA table and factor contribution

Source	DF	SS	MS	F	P	Contribution
A	1	44.925	44.925	101.85	0.001	51.22%
B	1	20.615	20.615	46.74	0.002	23.23%
C	1	19.547	19.547	44.32	0.003	22.00%
Error	4	1.764	0.441			3.55%
Total	7	86.851				100.00%

B. The employed data and data preprocess

This study uses a real sentiment data set from real world bloggers' comments. Table 4 introduces the brief background of the employed sentiment data. The data set comes from "ReviewCenter" (www.reviewcentre.com). By focusing on "Mobile phone (Phone)" related products comments, we collect 600 bloggers' comments. There are 400 positive and 200 negative comments in this imbalanced data set. In addition, because these bloggers' evaluations have no sentiment information, we use the 5-star rating system in this website to define bloggers' sentiments. A comment will be labelled as positive (negative) if the rate is above 4-stars (below 2-stars). Those comments whose rate is 3-stars have been disregarded. Moreover, five-fold cross validation experiment has been implemented in this study. By the way, some frequently used stop words should be removed. Readers can find a useful stop word listed at http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words. And, the package software QDA miner has been utilized to extract key words and construct TDM in this work. We use uni-gram to denote features. Each comment is converted into a vector of terms (keywords) with TF-IDF weights.

Table 4 The employed text sentiment data

No. of attributes	Data size	Class distribution
Fold #1: 2847	600	Positive: 400 Negative: 200
Fold #2: 2768		
Fold #3: 2922		
Fold #4: 2915		
Fold #5: 2941		

C. The Results

Traditionally, the easiest way to evaluate the classification performance is based on the confusion matrix shown as Table 5.

Table 5 Confusion matrix for binary class problem

	Predicted Positive	Predicted Negative
Actual Positive	*TP* (the number of True Positive)	*FN* (the number of False Negative)
Actual Negative	*FP* (the number of False Positive)	*TN* (the number of True Negative)

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

$$PA = TP/(TP + FN) \quad (9)$$

$$NA = TN/(FP + TN) \quad (10)$$

We use an integrated index, G-mean which is defined as equation (11) to measure the performance of imbalance classification. This measure 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.

$$G - mean = \sqrt{PA \times NA} \quad (11)$$

Table 6 lists results of several traditional methods without implementing feature selection. We can find original SVM that doesn't implementing any techniques for imbalanced data. It cannot identify any negative comments. Among three techniques, cost-adjusting method has the best performance 62.3%, then oversampling (45.7%) and under-sampling (27.4%). These results can be used as baseline for comparison.

Table 6 Summary of results of several traditional methods

	Original SVM		**Over-sampling**		**Under-sampling**		**Cost-adjusting**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
PA	100.0	0.0	22.3	9.7	8.0	3.8	73.8	32.4
NA	0.0	0.0	97.5	4.3	98.0	3.3	66.0	35.2
GM	0.0	0.0	45.7	9.5	27.4	6.3	62.3	22.5

Table 7 summarizes the comparison between our method and Sign-IG. Compared with Sign-IG, our method equally selects features from both positive and negative sentiments. In this table, we reduce dimension size to 50%, 25%, 10%, and 5% of original feature space. We can find our method outperforms Sign-IG in 50% and 25%. When the dimension size reduced to 10%, our method is almost equal to Sign-IG, but we have the smaller standard deviation. In 5%, Sign-IG is better than our method.

Table 7 Results of feature selection

Dimension size / Methods		**50%**		**25%**		**10%**		**5%**	
		Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
SignIG	**PA**	100.0	0.0	100.0	0.0	96.5	5.3	93.5	7.4
	NA	0.0	0.0	1.0	1.4	33.5	14.6	51.5	11.1
	GM	0.0	0.0	6.3	8.7	55.6	10.3	68.8	5.1
Our method	**PA**	100.0	0.0	99.8	0.6	96.8	3.0	95.0	4.3
	NA	0.5	1.1	2.5	2.5	32.0	7.6	43.0	7.2
	GM	3.2	7.1	12.1	11.3	55.3	6.3	63.7	4.2

Table 8 Results of feature selection methods integrated with cost-adjusting

Dimension size / Methods		**50%**		**25%**		**10%**		**5%**	
SignIG	**Cost**	1.6		1.6		1.4		1.4	
	PA	92.8	10.02	81.3	17.3	86.5	11.9	84.5	10.92
	NA	49	21.26	79	9.45	75	7.07	76.5	6.52
	GM	65.3	13.23	79.6	10.2	80.3	6.14	80.2	6.02
Our method	**Cost**	1.8		1.8		1.6		2	
	PA	74.3	17.69	85.3	9.33	90	8.84	81.5	8.02
	NA	86	9.78	83.5	5.18	75	4.68	84.5	6.94
	GM	79	6.98	84.2	4.87	82	3.76	82.9	6.6

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

According to results in Table 6, since cost-adjusting method is the best method for handling imbalanced sentiment data, we integrated our method and Sign-IG into cost-adjusting. Table 8 provides the results. From table 8, we can find our method is superior to traditional Sign-IG in all dimensions.

V. Conclusions

In this study, we have identified the key factors of imbalanced sentiment classification by using Taguchi method. They key factors are "data type (sentiment and non-sentiment)", "imbalance ratio (skewed distribution of collected text data)", and "dimension size (the amount of features)". Then, based these discovered key factors, we proposed a new feature selection method which equally selects features from both positive and negative sentiments to improve the performance of imbalanced sentiment classification. One case study from real world blogs has been provided to illustrate the effectiveness of our proposed approach.

After comparing several techniques for imbalanced data, we integrated cost-adjusting method into our method and Sign-IG. Experimental results indicated that the proposed method can improve imbalanced sentiment classification performance. To confirm the mentioned above results, additional experiments of other sentiment data should be implemented in the future.

References

[1] Channel Advisor Corporation, "Through the Eyes of the Consumer: 2010 Consumer Shopping Habits Survey," available: http://www.channeladvisor.com/, 2010.

[2] L. S. Chen, C. H. Liu, and H. J. Chiu, "A neural network based approach for sentiment classification in the blogosphere," *Journal of Informetrics*, 5(2): 313-322, 2011.

[3] K. Denecke and W. Nejdl, "How valuable is medical social media data? Content analysis of the medical web," *Information Sciences*, 179: 1870-1880, 2009.

[4] Y. M. Huang, T. C. Huang, and Y. M. Huang, "Applying an intelligent notification mechanism to blogging systems utilizing a genetic-based information retrieval approach," *Expert Systems with Applications*, 37: 705-715, 2010.

[5] V. Kumar, R. Venkatesan, and W. Reinartz, "Knowing what to sell, when, and to whom," *Harvard Business Review*, 131-137, 2006.

[6] J. Lee, D. H. Park, and I. Han, "The effect of negative online consumer reviews on product attitude: An information processing view," *Electronic Commerce Research and Applications*, 7: 341-352, 2008.

[7] Lightspeed research, "Consumer Reviews and Research online," http://www.lightspeedresearch.com/press-releases/consumers-rely-on-online-reviews-and-price-comparisons-to-make-purchase-decisions/, 2011.

[8] Nielsen Wire, "Social networks/blogs now account for one in every four and a half minutes online," available: http://blog.nielsen.com/nielsenwire/global/social-media-accounts-for-22-percent-of-time-online/, 2010.

[9] A. Rosenbloom, "The blogosphere," *Communications of the ACM*, 47 (12): 32–35, 2004.

[10] F. Wood-Black and T. Pasquarelli, "Blogs," *Journal of Chemical Health & Safety*, 14(2), pp. 37, 2007.

[11] W. Zhang, C. Yu, and W. Meng, "Opinion retrieval from blogs," *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 831-840, 2007.

[12] L. Zhu, A. Sun, and B. Choi, "Detecting spam blogs from blog search results," *Information Processing and Management*, 47(2):246-262., 2011.

[13] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: feature selection for opinion classification in web forums," *ACM Transactions on Information Systems*, 26(3), pp. 12:1–12:34 2008.

[14] A. Abbasi, H. Chen, S. Thoms, and T. Fu, "Affect analysis of web forums and blogs using correlation ensembles," *IEEE Transactions on Knowledge and Data Engineering*, 20(9):1168-1180, 2008.

[15] B. Li, S. Xu, and J. Zhang, "Enhancing clustering blog documents by utilizing author/reader comments," *In Proceedings of the 45th Annual Southeast Regional Conference*, 94-99, 2007.

[16] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," *In Proceedings of the 14th International Conference on World Wide Web*, 342-351, 2005.

[17] C. Whitelaw, N. Garg, and S. Argamon, "Using appraisal groups for sentiment analysis," *In proceedings of the ACM 14th Conference on Information and Knowledge Management*, 625-631, 2005.

[18] C. H. Wu, 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):165-182, 2006.

[19] H. Tang, S. Tan, and X. Cheng, "A survey on sentiment detection of reviews," *Expert Systems with Applications*, 36(7):10760-10773, 2009.

[20] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: opinion extraction and semantic classification of product reviews," *In Proceedings of the 12th International Conference on World Wide Web*, 519-528, 2003.

[21] P. Chaovalit and L. Zhou, "Movie review mining: A comparison between supervised and unsupervised classification approaches," *In Proceedings of the 38th Hawaii International Conference on System Sciences*, 2005.

[22] Q. Ye, Z. Zhang, and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches," *Expert Systems with Applications*, 36(3):6527-6535, 2009.

[23] S. Tan and J. Zhang, "An empirical study of sentiment analysis for Chinese documents," *Expert Systems with Applications*, 34(4):2622-2629, 2008.

[24] S. Wang, D. Li, X. Song, Y. Wei, and H. Li, "A feature selection method based on improved fisher's discriminant ratio for text sentiment classification," *Expert Systems with Applications*, 38(7):8696-8702, 2011.

[25] T. O'Keefe and I. Koprinska, "Feature selection and weighting methods in sentiment analysis," *In proceedings of the 14th Australasian Document Computing Symposium*, 2009.

[26] Z. Zheng, X. Wu, and R. Srihari, "Feature Selection for Text Categorization on Imbalanced Data," *ACM SIGKDD Explorations Newsletter*, 6(1):80-89, 2004.

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

[28] J. C. Na, C. Khoo, and P. H. J. Wu, "Use of negation phrases in automatic sentiment classification of product reviews," *Library Collections, Acquisitions, & Technical Services*, 29(2):180-191, 2005.

[29] M. Gamon, "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis," *In Proceedings of the 20th International Conference on Computational Linguistics*, 2004.

[30] F. Keshtkar and D. Inkpen, "Using sentiment orientation features for mood classification in blogs," *IEEE International Conference on Natural Language Processing and Knowledge Engineering*, 2009.

[31] M. Simeon and R. Hilderman, "Categorical Proportional Difference: A Feature Selection Method for Text Categorization," *In Proceedings of the 17th Australasian Data Mining Conference*, 2008.

[32] V. N. Vapnik, "The Nature of Statistical Learning Theory," *Springer-Verlag*, 1995.

[33] C. C. Chang and C. J. Lin, "LIBSVM: a Library for Support Vector Machines," Software, available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[34] J. Z. Zhang, J. C. Chen, and E. D. Kirby, "Surface roughness optimization in an end-milling operation using the Taguchi design method," *Journal of Materials Processing Technology*, 184, 233–239, 2007.

[35] C. W. Hong, "Using the Taguchi method for effective market segmentation," *Expert Systems with Applications*, doi:10.1016/j.eswa.2011.11.040, 2011.

[36] W. T. Chien and C. S. Tsai, "The investigation on the prediction of tool wear and the determination of optimum cutting conditions in machining 17-4PH stainless steel." *Journal of Materials Processing Technology*, 140, 340–345, 2003.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)