International Journal of Applied Science and Engineering

Special issue: The 10th International Conference on Awareness Science and Technology (iCAST 2019)

New term weighting methods for classifying textual sentiment data

Jing-Rong Chang, Long-Sheng Chen*, Chia-Wei Chang

Department of Information Management, Chaoyang University of Technology, Taichung, Taiwan, R.O.C.

ABSTRACT

In current society, people can easily use social media to express their own opinions toward products and services. These online comments can influence other customers' purchase behaviors. Especially those negative reviews and comments can hurt the images of companies. Consequently, to identify the sentiment of social media users from a large amount comments is one of crucial issues. In recent years, machine learning approaches have been considered as one of possible solutions for recognizing sentiment of text reviews. But, when using these methods to sentiment classification, traditional term weighting methods including Term Presence (TP), Term Frequency (TF), and Term Frequency-Inverse Document Frequency (TF-IDF) often have been utilized for describing the collected textual reviews. However, those conventional term weighting methods cannot have positive effect on improving the classification performance of text sentiment data. Therefore, this study aims to propose two new term weighting methods called Categorical Difference Weights (CDW) and TF-CDW by integrating class information into term weights of textual data to construct Term-Document Matrix (TDM). Then, Support Vector Machines (SVM) will be employed to build classifiers. Finally, we will use several actual cases to demonstrate the effectiveness of our presented methods. Compared to traditional term weighting methods, results showed that our methods indeed outperform TF, TP and TF-IDF.

Keywords: Sentiment classification; Term weighting; Class information; Text mining; Product reviews.

OPEN ACCESS

Received: April 14, 2020
Revised: May 7, 2020
Accepted: July 26, 2020

Corresponding Author:
Long-Sheng Chen
lschen@cyut.edu.tw

Publisher:
Chaoyang University of Technology
ISSN: 1727-2394 (Print)
ISSN: 1727-7841 (Online)

1. INTRODUCTION

With the quick development of social media such as Facebook, twitters and so on, the number of personal reviews and opinions of Internet users increase remarkably (Chatterjee and Kar, 2020). Customers can easily express their feelings and usage experiences regarding the purchased products and experienced services. These personal opinions, especially negative comments, might have a significant influence on other consumers' purchasing intensions. For instance, many people might learn how others' viewpoints of a specific product before buying (Zhang et al., 2008; Mekawie and Hany, 2019; Xu et al., 2020). An enterprise can improve their service and product quality based on customer's reviews no matter they are positive or negative. However, some negative product evaluations which often spread very quickly could reduce consumers' purchase intentions. Therefore, to effectively recognize the sentiment of consumers from a large amount of online textual reviews had become one of critical issues.

In recent years, sentiment classification which classifies textual sentiment data into positive or negative group has attracted lots of attention (Liu et al., 2005; Zhao et al., 2020; Kong et al., 2020). Generally speaking, sentiment classification aims to recognize type of sentiment of reviews from customers' opinions for certain products or services

https://doi.org/10.6703/IJASE.202009_17(3).257

(Ye et al., 2009; Mekawie and Hany, 2019). Lots of works have been presented to solve sentiment classification problems for textual data. These studies could be grouped into two categories (Tan and Zhang, 2008; Akhtar et al., 2020). The first group mainly contains machine learning techniques. The methods in this group try to build classifiers based on labeled textual reviews and then recognize the sentiment of new coming reviews based on the built classifier. The second group mainly includes semantic orientation approaches. They classify features into two classes (positive or negative), and then count the overall positive and negative scores in the examples to determine the sentiment of review.

From available published works, methods in the first group have been considered as one of useful solutions for classifying sentiment. For instances, Na et al. (2005) applied POS (Part of Speech) tags based negation phrases with Support Vector Machines (SVM) to enhance the classification performance of textual reviews. In the work of Ye et al. (2009), they applied machine learning methods including Naive Bayes (NB), SVM, and the character based N-gram model, to classify the sentiment of online travel reviews. In the study of Tan and Zhang (2008), they compared four feature selection techniques and five machine learning approaches on classifying Chinese sentiment classification. In the work of Bai (2010), a heuristic search-enhanced Markov blanket model has been presented. In his model, SVM has been employed to predict consumer sentiments from online reviews. Akhtar et al. (2020) presented a multi-task learning framework for classifying sentiment. In the work of Gokalp et al. (2020), a new wrapper feature selection algorithm based on Iterated Greedy (IG) metaheuristic has been presented for sentiment classification. For large-scale and multi-domain e-commerce platform product review, Xu et al. (2020) presented a Naïve Bayes (NB) learning framework for identifying sentiment.

When applying machine learning techniques to textual data. The textual data would be represented by a feature vector which is built by calculating the weights of features (terms) in the documents and then build a term-document matrix (TDM). In a TDM, documents could be represented by vectors which are expected to indicate as much information of the documents as possible (Tian and Tong, 2010). To accurately and efficiently represent collected texts, the term weighting method plays an important role in the process of build TDM (Tian and Tong, 2010). In related works of text classification, lots term weight methods, such as term frequency-inverse document frequency (TF-IDF), term frequency (TF), inverse document frequency (IDF), and term presence (TP), and so on, have been successfully developed. However, these conventional methods cannot have positive and direct impact on the improvement of sentiment classification performances. They are calculated by the terms' occurrence frequency in a document. Or they are represented to show if a term appears in a document or not. Therefore, this study will propose new term weighting methods called Categorical Difference Weights (CDW) and TF-CDW by introducing class information while counting the weight of a term in a document. Moreover, the most common and easiest dimension reduction method, feature frequency (FF) technique, will be employed to study the effect of CDW and TF-CDW on feature selection. Then, we use SVM to construct classifiers for identifying textual sentiment data. Finally, several actual cases of online product reviews will be provided to illustrate the effectiveness of our proposed CDW and TF-CDW methods.

2. RELATED WORKS

2.1 Sentiment Classification

A brief literature review regarding sentiment classification will be given in this section. Sentiment analysis is becoming more and more important as the number of digital text resources (Gokalp et al., 2020). Table 1 summarizes and analyzes available related works. In this table, we can find that there are three groups of methods, SVM, NB, and others (N-gram, Maximum entropy, decision trees, etc.) mainly applied to sentiment classification domains. Among them, SVM is the most popular method. Most of listed researches indicate that SVM outperforms other methods. This is the reason why we employ SVM in this work.

Considering term weighting approaches that aim to indicate the significant of a term in a document (Aizawa, 2003), TF, TF-IDF, and TP are the most common used to count the weight of a term (Na et al., 2005; Pang et al., 2002; Martineau and Finin, 2009; O'Keefe and Koprinska, 2009). TF denotes the occurrence frequency of a term occurs in review, and TF-IDF is the combination of TF and IDF weights. IDF represents the general importance of a term in overall reviews. IDF and TF-IDF have be defined in Equations (1) and (2).

$$IDF = \log \frac{\textit{The number of total documents}}{\textit{The number of documents include a term}} \tag{1}$$

$$TF - IDF = TF \times IDF \tag{2}$$

If one term's TF-IDF weight is high, it represents this term occurs frequently and only appears in the part of overall reviews.

TP has been first used to represent term weights by Pang et al. (2002) in sentiment analysis. TP is very like to TF, except that rather than using the frequency of a unigram as its value. In TP, we only use binary presentation "1" and "0" to denote that a term exists or not in the review (O'Keefe and Koprinska, 2009). To sum up, TP (binary weights) has the value 1 if the term exists in the review, 0 means absence (Na et al., 2005).

In both the works of Pang et al. (2002) and O'Keefe and Koprinska (2009), they utilized SVM to classify movie reviews, and their experimental results indicated that using

Table 1. Summary of related works in sentiment classification.

	Machine learning methods			Term weighting methods			Employed data		
	SVM	NB	Others	TF	TF-IDF	TP	D1	D2	D3
Pang et al. (2002)	✓	✓	✓	✓		✓	✓		
Chaovalit and Zhou (2005)			✓	-	-	-	✓		
Na et al. (2005)	✓			✓	✓	✓	✓		
Whitelaw et al. (2005)	✓			✓			✓		
Kennedy and Inkpen (2006)	✓					✓	✓		
Abbasi et al. (2007)	✓					✓	✓	✓	
Li et al. (2007)	✓		✓			✓	✓		
Tan and Zhang (2008)	✓	✓	✓		✓		✓		
Yu et al. (2008)	✓			-	-	-			✓
Zhang et al. (2008)	✓	✓	✓		✓		✓		
Chen and Chiu (2009)			✓		✓		✓		
Martineau and Finin (2009)	✓				✓		✓		
O'Keefe and Koprinska (2009)	✓	✓		✓	✓	✓	✓		
Ye et al. (2009)	✓	✓	✓	✓			✓		
Bai (2010)	✓	✓	✓			✓	✓		
Zhang et al. (2011)	✓				✓				✓

Note: D1, D2, and D3 represent "product comments & review", "web forum postings", and "others (political issues and so on)", respectively.

TP can have a better performance than using TF or TF-IDF. But, in the work of Na et al. (2005), they employed SVM to identify sentiment of product reviews, and they concluded that TF-IDF has the best performance (the next place is TP and then TF) That's also why TP and TF-IDF have been utilized in most of works listed in Table 1. In this study, we compared our proposed CDW and TF-CDW method with TP, TF-IDF, and TF weights. Additionally, product comments and reviews are the most popular employed data type.

2.2 Dimension Reduction Methods

With increasing of the textual data in cyberspace, how to extract important information from a huge amount of reviews in social media websites have been become one of critical problems. Feature selection in text mining is to extract key terms for representing all collected documents (Wang et al., 2010; Gokalp et al., 2020; Akhtar et al., 2020). It aims to reduce feature space, shorten computational costs, remove noises, and improve the classification performance. (Chen et al., 2009; Li et al., 2007; Polat and Gunes, 2009; Karabatak and Ince, 2009; Kong et al., 2020).

Lots of feature selection methods have been developed for dimension reduction, such as mean TF-IDF (Tang et al., 2005) and feature frequency (FF). Among them, FF is the most common and easiest technique for selecting relevant terms in the documents. According to available published literatures (Na et al., 2005; Pang et al., 2002), feature frequency based unigrams have been proven that it could have good performances. For example, in the experiments of Na et al. (2005), they indicated that feature frequency based unigrams out-performed terms labeled with POS (part of speech) tags. Pang et al. (2002) also verified that only using unigrams as features are better than bigrams, combinations of unigrams and bigrams, and POS tags. Consequently, this study employs unigram to select attributes of textual data.

2.3 Support Vector Machines

SVM developed by Vapnik (1995) is a supervised machine learning technique based on risk minimization principle of statistical learning theory. In sentiment classification area, SVM has been successfully applied to solve classification problems by finding a hyperplane of maximal margin. Lots of studies (Ye et al., 2009; Tan and Zhang, 2008; Na et al., 2005; Pang et al., 2002; Martineau and Finin, 2009; O'Keefe and Koprinska, 2009; Li et al., 2007; Zhang et al., 2019) reported that SVM had a superior performance on sentiment classification. Moreover, Khan et al. (2009) also surveyed 336 research papers which have used machine learning approaches. The reported results indicated that SVM has been one of most widely used machine learning methods. Additionally, SVM has several advantages including the use of kernels (no need to acknowledge the non-linear mapping function), the absence of local minima (quadratic problem), the sparseness of solution and the generalization capability obtained by optimizing the margin (Cerqueira et al., 2008; Bansal and Srivastava, 2018).

In fact, SVM builds a decision boundary between two classes by mapping the training examples onto a higher dimensional space through a kernel function, and then finds a maximal margin hyperplane within that space. Finally, this hyperplane can be viewed as a classifier. A brief introduction of SVM operations can be found in the following.

Giving n examples $S = \{x_i, y_i\}_{i=1}^{n}$, $y_i \in \{-1, +1\}$, where x_i represents the attributes; y_i is the class; and i is the number of training examples. The decision hyperplane of SVM can be defined as (w, b), where w is a weight vector and b a bias. Let w_0 and b_0 denote the optimal values of the weight vector and bias. Correspondingly, the optimal hyperplane can be formatted as Equation (3).

$$w_0^T x + b_0 = 0 \tag{3}$$

In order to find the optimum values of w and b, it has to solve the following optimization problem.

$$\begin{aligned} &\min_{w,b,\xi} && \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \\ &\text{Subject to} && y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i \\ & && \xi_i \geq 0 \end{aligned} \tag{4}$$

where ξ is the slack variable, C is the user-specified penalty parameter of the error term ($C > 0$), and ϕ is the kernel function. In this work, the common used radial basis function (RBF) (Hsu et al., 2006) kernel function has utilized and it could be defined in Equations (5).

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\},\ \gamma > 0 \tag{5}$$

where, γ is kernel parameter and are user-defined.

3. PROSED METDOLOGY

3.1 The Proposed CDW and TF-CDW Approach

In this section, we will demonstrate how to implement our proposed CDW and TF-CDW methods. Actually, our CDW and TF-CDW is extended from Categorical Proportional Difference (CPD) (Simeon and Hilderman, 2008) which is originally developed as a quick term selection method for text classification. Later, the work of O'Keefe and Koprinska (2009), they successfully used CPD to sentiment classification. Consequently, first, we should introduce CPD which has been shown in Equation (6).

$$CPD = \frac{|PositiveDF - NegativeDF|}{PositiveDF + NegativeDF} \tag{6}$$

According to the definition of CPD, we can find that CPD compute the positive document frequency (Positive DF) and negative document frequency (Negative DF), which means the number of documents including a term in positive class, of a term separately, and then calculate the proportional difference of a term in both classes.

In Equation (6), we can understand that the score of CPD will be in [0, 1] interval. If one term only appears in positive document or negative document, we can find the CPD score is equal to 1. Then, this term will be considered as important for classification. On the contrary, if one term equally appears in positive and negative documents, its value of CPD will be equal to 0. And this term will be viewed as unimportant. From previous works, we can know CPD could discover the crucial attributes by introducing class information. But, CPD has a serious drawback that we cannot select important attributes when lots of attributes have the same CPD scores. It's especially true when we reduce feature space to a small dimension size. To clearly illustrate this disadvantage, we take Table 2 for example.

In this table, assume we need to select important attributes (terms) from six candidates (Attributes 1-6) based on the rank of their CPD scores. In this example, we can find that attribute 1 is more important than attribute 2-4, if we only consider occurrence frequency. But, all of them have the same CPD score. Under such situation, we cannot know how to select important attributes because CPD cannot efficiently indicate terms' significances. Consequently, to improve CPD, we modified CPD and proposed Enhanced CPD (ECPD).

Equations (7)-(9) defines ECPD in different conditions. In different conditions, the considerations and ECPD Equation have listed as follows.

Condition #1

When a term's 'Positive DF' or 'Negative DF' is equal to zero, or 'Positive DF' is equal to 'Negative DF', ECPD can be defined as Equation (7).

$$ECPD = |PositiveDF - NegativeDF| \tag{7}$$

Condition #2

When a term's 'Positive DF' is greater than 'Negative DF', ECPD can be defined as Equation (8).

$$ECPD = \frac{PositiveDF}{NegativeDF} \tag{8}$$

Condition #3

When a term's 'Positive DF' is less than 'Negative DF', ECPD can be defined as Equation (9).

$$ECPD = \frac{NegativeDF}{PositiveDF} \tag{9}$$

In order to testify that our proposed ECPD indeed can enhance CPD method, let's go back the illustrative examples in Table 2. In Table 2, originally we cannot identify the importance of attributes 1-4. But, by using ECPD, we can easily rank the importance of attribute 1-4. So, our proposed ECPD can indicate a term's importance in the documents compared with conventional CPD.

Next, based on ECPD, we present a new term weighting method called Categorical Difference Weights (CDW) which has been defined as Equation (10).

Table 2. An illustrative example for CPD and ECPD

	Positive DF	Negative DF	CPD	ECPD
Attribute 1	20	0	1	20
Attribute 2	0	15	1	15
Attribute 3	10	0	1	10
Attribute 4	0	2	1	2
Attribute 5	3	2	0.2	1.5
Attribute 6	1	1	0	0

$$CDW = \frac{ECPD}{PositiveDF + NegativeDF} \tag{10}$$

In CDW, we view a term's ECPD score as its weight in overall documents. In addition, Equation (10) has been multiplied by TF, and then another weighting method, TF-CDW, can be defined as Equation (11).

$$TF - CDW = TF \times \frac{ECPD}{PositiveDF + NegativeDF} \tag{11}$$

3.2 Implemental Procedure

Next, we will illustrate the detailed implemental procedure listed in Fig. 1. In fact, the implemental algorithm can be divided into four steps. They are described as follows.

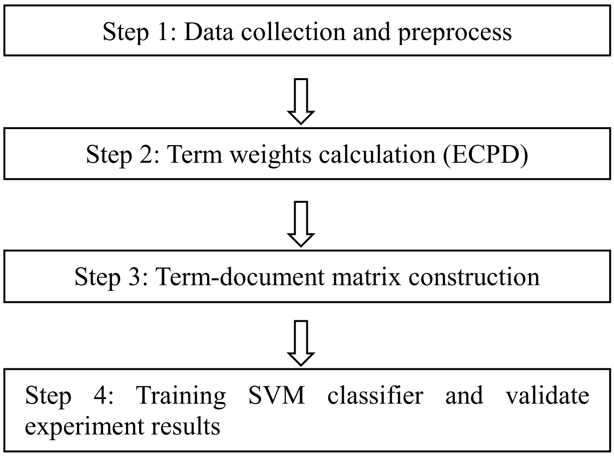
Step 1: Data collection and preprocess
Step 2: Term weights calculation (ECPD)
Step 3: Term-document matrix construction
Step 4: Training SVM classifier and validate experiment results

Fig. 1. The implemental procedure of the experiment.

Step 1: Data collection and preprocess

In this step, we will collect textual comments or reviews regarding products or services to be our experiment corpuses. In addition, to process collected textual sentiment data, we use unigram to denote attributes and segment sentences. Not all the words in sentences are useful for classifying semantic orientations or related tasks. Accordingly, we will remove stop words and irrelevant words. Some words with low occurrence frequency will also be removed. Finally, a set of candidate terms will be constructed for next step.

Step 2: Term weights calculation (ECPD)

In this step, we calculate ECPD scores according to Equations (7)-(9) for all candidate terms. And a term's ECPD score could be viewed as its weights in documents. Then, we can use this ECPD weights to construct TDM. Take Table 2 for example. Based on "Positive DF" and "Negative DF" of every single term, we can compute ECPD scores. The ECPD scores of attributes 1-6 are, 20, 15, 10, 2, 1.5, and 0, respectively.

Step 3: Term-document matrix construction and dimension reduction

In this step, the major task is to construct TDM with CDW and TF-CDW weights. According to the calculated ECPD scores in step 2, we can compute the CDW and TF-CDW weights by using Equations (10) and (11). Use the same case in Table 2, the CDW weights of attributes 1-6 are $\frac{20}{20}$, $\frac{15}{15}$, $\frac{10}{10}$, $\frac{2}{2}$, $\frac{1.5}{5}$, and $\frac{0}{2}$ (i.e. 1, 1, 1, 1, 0.3, 0), individually.

Then, Fig. 2 shows the operations of constructing CDW and TF-CDW matrix. In part (a) of this Figure, it shows an original TDM with TF. In this case, we have six candidate attributes (1-6) and five documents. Part (b) shows the TDM with CDW weights. Compared with part (a), CDW weights have replaced TF. Next, the CDW weights in part (b) have been multiplied by TF described in part (a). Finally, we can get a TDM with TF-CDW shown in part (c).

Another task of this step is to reduce feature space for shorten training time of SVM. In this work, we implement two kinds of experiments. The first one is merely to testify the effects of introducing CDW and TF-CDW weights. The second one is to discover the proposed methods' influences in feature selection.

(a) TDM with TF

Attributes Documents	1	2	3	4	5	6	Class
Doc #1	**4**	0	0	0	**1**	0	+1
Doc #2	**1**	0	**3**	0	0	**1**	+1
Doc #3	0	**3**	0	0	0	0	-1
Doc #4	0	**2**	0	**2**	**2**	**1**	-1
Doc #5	0	0	0	**1**	0	0	-1

(b) TDM with CDW

Attributes Documents	1	2	3	4	5	6	Class
Doc #1	**1**	0	0	0	**0.3**	0	+1
Doc #2	**1**	0	**1**	0	0	**0**	+1
Doc #3	0	**1**	0	0	0	0	-1
Doc #4	0	**1**	0	**1**	**0.3**	**0**	-1
Doc #5	0	0	0	**1**	0	0	-1

(c) TDM with TF-CDW

Attributes Documents	1	2	3	4	5	6	Class
Doc #1	**4**	0	0	0	**0.3**	0	+1
Doc #2	**1**	0	**3**	0	0	**0**	+1
Doc #3	0	**3**	0	0	0	0	-1
Doc #4	0	**2**	0	**2**	**0.6**	**0**	-1
Doc #5	0	0	0	**1**	0	0	-1

Fig. 2. An illustrative example for CDW and TF-CDW

In this study, we use FF which ranks term frequency to select key attributes to reduce dimension size. To have a comparison base, we arbitrarily determine six size of feature space including 1000, 700, 400, 200, 100, and 50. These selected important terms can be used to denote our collected corpuses. After determining the dimension size, we compute the term weights. Each text review could be described in 5 kinds of weights, including CDW, TF-CDW, TP, TF, and TF-IDF. Furthermore, a 5-fold cross validation experiment has been used in this study for building training and test data sets.

Step 4: Training SVM classifier and validate experiment results

In the last step, the training data sets will be utilized to build SVM classifiers, and then input the remaining test data set to validate the constructed SVM models. In addition, our CDW, TF-CDW methods will be compared with traditional TP, TF, and TF-IDF methods. Finally, based on the results, we can make some concluding remarks.

4. IMPLEMENTATION

4.1 Data Collection and Data Preprocess

In this study, we use three sentiment textual data sets including two real cases of product reviews and one famous movie reviews database to evaluate the effectiveness of the proposed weighting methods. Table 3 gives a brief introduction of the employed textual sentiment data. In this table, the first data set is from movie reviews (Movie) database. We merely randomly select 400 comments involving 200 positive and 200 negative comments to be our experimental corpus.

The second and third data sets are collected in “Review Centre” website which contains millions of consumers’ product reviews. Those provided reviews database have been categorized into some groups. In this study, we merely select of “MP3 player (MP3)” and “Mobile phone (Phone)” topics related product reviews. Next, we manually preprocess those collected textual product reviews. Then, QDA miner package software will be employed to handle those prepared text reviews and transform them into TDM. In addition, the 5-star rating system of “Review Centre” has been utilized to define its class labels. If one review has 4 or more stars (2 or less stars), this review will be labeled as positive (negative) sentiment. Besides, the comments that have 3 stars will be ignored. After preparation process, 400 and 800 product reviews are collected from MP3 and Phone, respectively.

Table 3. The employed three textual sentiment data sets

No	Data Set	Notation	Source	No. of attributes	Data Size	Class distribution
1	Movie Review	Movie	http://www.cs.cornell.edu/people/pabo/movie-review-data/	4428	400	Positive: 200 Negative: 200
2	MP3 product evaluation	MP3	http://www.reviewcentre.com	1382	400	Positive: 200 Negative: 200
3	Cellular phone review	Phone	http://www.reviewcentre.com	2323	800	Positive: 400 Negative: 400

Table 4. Results of SVM in three corpuses without implementing feature selection method.

Weights	TP		TF		TF-IDF		CDW		TF-CDW	
Corpuses (dimension size)	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)
Movie (4428)	76.00	5.11	71.25	3.19	75.75	4.20	92.75	2.05	87.75	3.79
MP3 (1382)	84.00	8.07	82.00	6.16	81.50	8.59	89.25	6.82	86.50	9.50
Phone (2323)	85.38	1.57	85.13	2.59	85.13	2.09	84.50	4.29	84.88	3.23

Note: “Mean” is the average value and “SD” represents standard deviation.

By the way, we use unigram to segment collected sentences. Some frequently used stop words and the words with low occurrence frequency have been removed. After data preprocessing phase, movie review data has 4428 attributes are left for advanced analysis. In MP3 and Phone data, 1382 and 2323 attributes have been extracted. Then, each comment is converted into a vector of terms (keywords) with TP, TF, TF-IDF, CDW, and TF-CDW weights. Moreover, LIBSVM developed by Chang and Lin (2001) has been utilized to build SVM classifier. The default kernel function is radial basis function (RBF). All optimal parameter settings of SVM could be obtained automatically in LIBSVM.

4.2 Experimental Results

4.2.1 Results without implementing dimension reduction

This section provides results of different weighting methods with SVM. We do not consider the influence of dimension reduction. After 5-fold cross validation experiment, the results including mean value (Mean) and standard deviation (SD) can be summarized in Table 4. First, we compare the difference between/among three conventional term weights. Results in all of three corpuses indicated that TP has better performance than TF and TF-IDF. However, as shown in H1-H2 of Table 5, the p-values of these two hypotheses are 0.185 and 0.324 (> 0.05). We cannot reject the null hypothesis (H_0). Consequently, we have 95% confidence to believe the differences between TP and TF, and TP and TF-IDF are not statistically significant.

Next, we will compare our CDW with TF-CDW. From Table 4, except the result in Phone data, CDW has better mean and smaller standard deviation (Movie: Mean = 92.75%, SD = 2.05%; MP3: Mean = 89.25%, SD = 6.82%) than TF-CDW (Movie: Mean = 87.75%, SD = 3.79%; MP3: Mean = 86.50%, SD = 9.50%). In order to confirm this conclusion in advance, we implement statistic hypothesis test listed in H3 of Table 5. As a result, the p-value is 0.126 (> 0.05). We cannot reject the null hypothesis (H_0). Therefore, we have 95% confidence to believe that the performance between TF-CDW and CDW is not significant. In spite of having no statistical evidence to claim CDW outperforms TF-CDW, the average classification accuracy of CDW is slightly better than TF-CDW according to results in Table 4.

We continue to testify if CDW or TF-CDW is better than traditional TP, TF, and TF-IDF or not. Hypotheses H4-H6 are presented to verify that CDW outperforms conventional TP, TF, and TF-IDF. As a result, all p-values of H4-H6 are all less than 0.05. That means that we can reject all null hypotheses. Therefore, we have 95% confidence to believe that our CDW is better than traditional weighting methods, TP, TF, and TF-IDF. The same conclusion also could be made from results of hypotheses H7-H9. Accordingly, we can say that our proposed CDW and TF-CDW are better than TP, TF, and TF-IDF without implementing feature selection techniques.

4.2.2 Results with implementing dimension reduction technique

This subsection will study the effects of different kinds of weights by feature selection approaches. This work employs the most common and easiest dimension reduction approach, feature frequency (FF), which is to rank terms by their term frequencies and then select key terms based on their ranks. The proposed CDW and TF-CDW will be compared with TP, TF, and TF-IDF under different reduced dimension sizes.

Table 6 and Fig. 3 show the results of Movie data. From Fig. 3, we can clearly find that CDW outperforms others when dimension size has been reduced from 4428 to 1000, 700, 400, and 200. All hypothesis tests, whose p-values are far less than 0.05, confirm the superiority of CDW. We have statistical evidence to believe CDW is better than others. But, when continuously downsizing the feature space to 100 and 50, from Table 6, we find TF-CDW has a better performance than others including CDW, but the results of hypotheses merely indicated that TF-CDW is better than TP at 95% confidence level.

Table 7 and Fig. 4 show the results of MP3 data. From Fig. 4, we can find that CDW outperforms others under all dimension sizes. But, from the results of hypothesis tests, we merely have statistical evidences to believe CDW is better than TF and TF-IDF, not including TF-CDW and TP. However, from the numerical results of Table 7, we also can conclude that CDW is slightly better than TF-CDW and TP, but far better than TF and TF-IDF.

Table 8 and Fig. 5 show the results of Phone data. From Fig. 5, when dimension size has been reduced from 4428 to 1000, 700, 400, and 200, it's hard to say which weighting method is better than another. But, in low dimension which feature size is equal to 100 and 50, the results of hypothesis tests indicated that CDW and TF-CDW have the best performance, respectively. Therefore, the results of Phone data only can tell us that CDW or TF-CDW could have better performance than traditional TP, TF, and TF-IDF in low dimension size (100 & 50). In other dimension sizes, it's hard to identify which one weighting method is better than the other.

Table 5. Hypothesis tests.

No	Hypothesis	P-value	Conclusion
H1	$H_0: \mu_{TP} \leq \mu_{TF}$ $H_1: \mu_{TP} > \mu_{TF}$	0.185	Don't reject H$_0$
H2	$H_0: \mu_{TP} \leq \mu_{TF-IDF}$ $H_1: \mu_{TP} > \mu_{TF-IDF}$	0.324	Don't reject H$_0$
H3	$H_0: \mu_{CDW} \leq \mu_{TF-CDW}$ $H_1: \mu_{CDW} > \mu_{TF-CDW}$	0.126	Don't reject H$_0$
H4	$H_0: \mu_{CDW} \leq \mu_{TP}$ $H_1: \mu_{CDW} > \mu_{TP}$	0.002	Reject H$_0$
H5	$H_0: \mu_{CDW} \leq \mu_{TF}$ $H_1: \mu_{CDW} > \mu_{TF}$	0.000	Reject H$_0$
H6	$H_0: \mu_{CDW} \leq \mu_{TF-IDF}$ $H_1: \mu_{CDW} > \mu_{TF-IDF}$	0.001	Reject H$_0$
H7	$H_0: \mu_{TF-CDW} \leq \mu_{TP}$ $H_1: \mu_{TF-CDW} > \mu_{TP}$	0.028	Reject H$_0$
H8	$H_0: \mu_{TF-CDW} \leq \mu_{TF}$ $H_1: \mu_{TF-CDW} > \mu_{TF}$	0.004	Reject H$_0$
H9	$H_0: \mu_{TF-CDW} \leq \mu_{TF-IDF}$ $H_1: \mu_{TF-CDW} > \mu_{TF-IDF}$	0.010	Reject H$_0$

Note: 95% confidence level has been used for doing hypothesis tests.

Table 6. Results of movie review for implementing dimension reduction approach.

Weights	TP		TF		TF-IDF		CDW		TF-CDW	
Dimensions	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)
1000	74.75	2.85	70.50	2.44	75.00	2.65	80.50	2.59	73.75	1.53
700	74.25	3.81	70.00	3.19	71.00	5.11	77.00	3.49	73.50	6.09
400	70.75	2.27	69.50	2.44	69.75	4.37	73.50	4.63	71.50	4.54
200	67.50	2.50	67.00	4.29	69.75	2.85	72.50	4.68	70.25	3.24
100	64.25	2.44	65.75	3.60	68.50	4.28	67.25	2.98	69.25	2.59
50	60.00	3.06	65.00	4.15	63.75	3.19	63.50	3.99	65.50	4.38

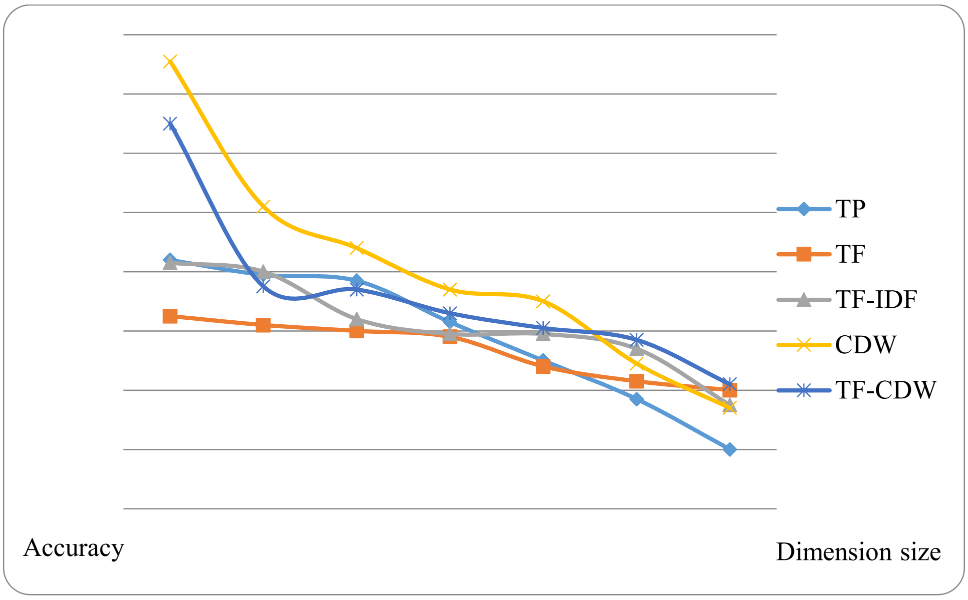

Fig. 3. Results of implementing dimension reduction approach (Movie).

Table 7. Results of MP3 review for implementing dimension reduction approach.

Weights	TP		TF		TF-IDF		CDW		TF-CDW	
Dimensions	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)
1000	84.75	7.36	82.25	6.15	79.75	8.72	87.50	7.71	84.00	8.36
700	86.50	7.78	81.50	5.96	79.00	7.15	88.00	7.48	85.50	8.82
400	86.00	7.20	81.50	6.27	81.00	6.58	87.75	7.04	86.00	8.50
200	81.50	5.18	79.50	5.63	81.75	5.05	85.50	3.38	85.50	6.03
100	79.75	5.69	75.50	7.74	78.75	7.07	80.75	8.73	79.00	6.21
50	76.50	5.89	74.75	8.26	75.50	5.84	77.75	6.75	76.00	4.28

Table 8. Results of phone review for implementing dimension reduction approach.

Weights	TP		TF		TF-IDF		CDW		TF-CDW	
Dimensions	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)
1000	84.50	1.73	84.50	2.59	85.75	2.70	85.13	3.91	85.50	4.13
700	84.25	2.36	83.25	2.44	84.88	1.90	84.25	4.18	84.63	4.67
400	83.00	2.23	83.50	2.15	81.25	2.34	83.25	4.04	84.25	2.81
200	82.75	2.05	81.63	2.71	81.13	1.73	82.75	3.30	82.38	3.23
100	79.38	1.53	78.88	3.63	79.00	1.30	82.00	1.95	81.00	2.15
50	74.63	2.36	74.63	3.02	75.13	2.18	77.13	3.21	79.13	2.75

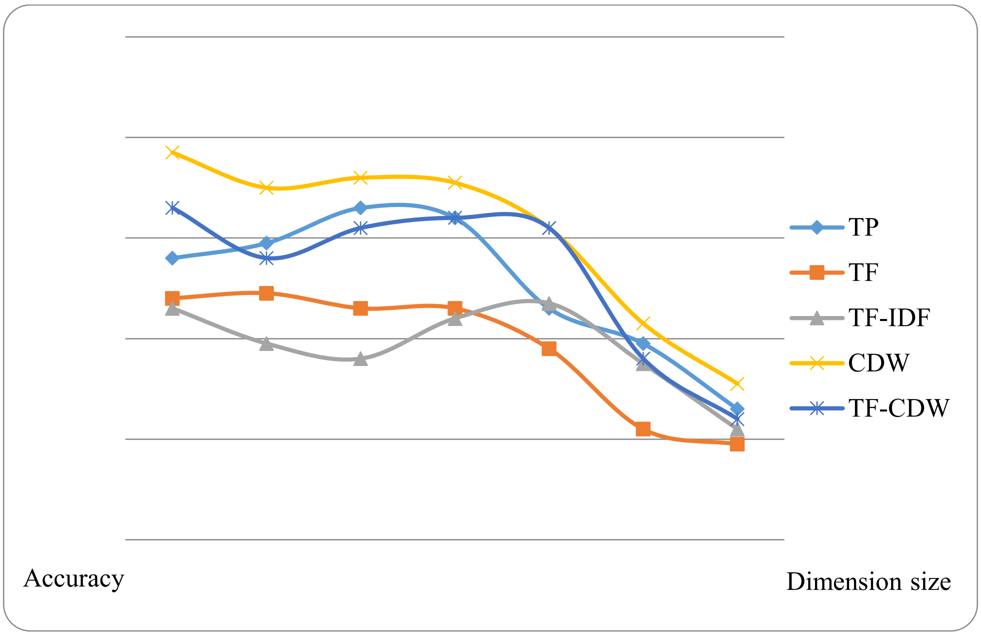

Fig. 4. Results of implementing dimension reduction approach (MP3).

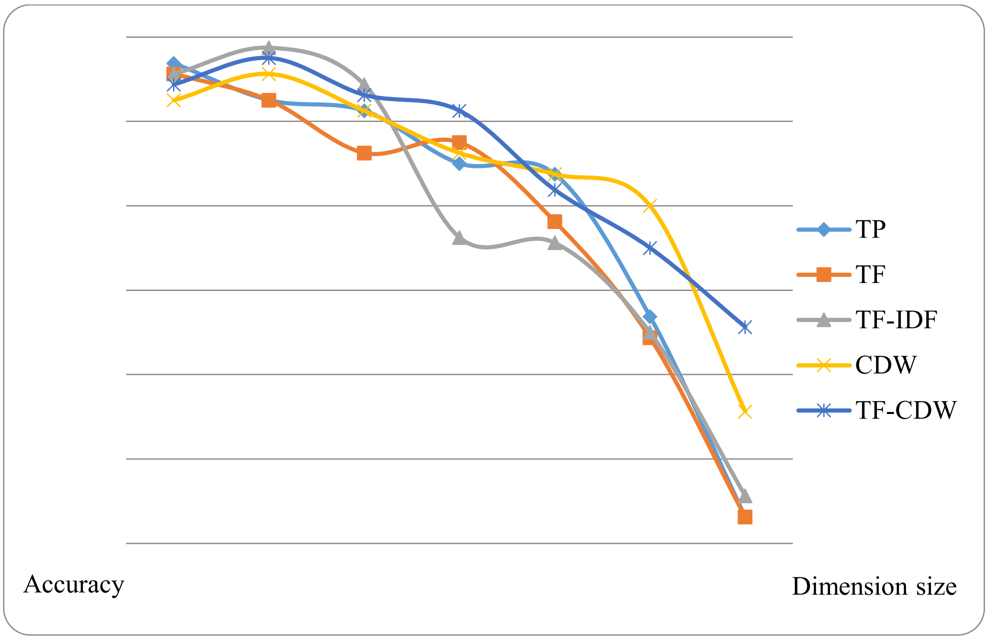

Fig. 5. Results of implementing dimension reduction approach (Phone).

5. CONCLUSIONS

By enhancing CPD, we proposed new term weighting methods, CDW and TF-CDW, to improve the sentiment classification performance of textual reviews in social media. Several real case of product reviews have been collected from social media websites. From experimental results, we could make some concluding remarks. Firstly, if we don't take feature reduction techniques into consideration, our CDW and TF-CDW outperform conventional weighting approaches, TP, TF and TF-IDF. Our methods can have better performances even in low dimension space. It means that one classifier combined with our method can save lots of computational source and keep the performance in the same time. When dealing with the increasing amount of text reviews, it's very important. Secondly, after implementing dimension reduction techniques, although the classification performance decreases eventually, the proposed CDW still outperforms TF and TF-IDF methods. But, the performance gaps between our CDW and traditional weighting methods will be shortened. Consequently, even considering dimension reduction, our proposed CDW method indeed can improve the performance of sentiment classification compared with conventional weights. That proves our weighting methods indeed can replace conventional weighting methods.

In addition, to avoid other uncontrolled factors, this study merely uses FF method to select features. But, there are lots of feature selection methods, such as Chi-square, information gain, mutual information and so on in text classification. Integrating these methods into our CDW might be potential direction of future works. Moreover, this study only uses movie, phone, MP3 product reviews and its data size is not huge enough. Therefore, to include more different kinds of product reviews or large size of data might be one of the potential directions of future works.

ACKNOWLEDGMENT

This work was financially supported in part by Ministry of Science and Technology, Taiwan (Grant No. MOST 108-2410-H-324-009). Authors express our appreciation to their funding, and to the useful suggestions from anonymous reviewers.

REFERENCES

Abbasi, A., Chen, H., Salem, A. 2007. Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. ACM Transactions on Information Systems, 26.

Aizawa, A. 2003. An information-theoretic perspective of TF-IDF measures. Information Processing and Management, 39, 45–65.

Akhtar, S., Garg, T., Asif Ekbal, A. 2020. Multi-task learning for aspect term extraction and aspect sentiment classification. Neurocomputing, in press.

Bai, X. 2010. Predicting consumer sentiments from online text. Decision Support Systems, doi:10.1016/j.dss.2010.08.024.

Bansal, B., Srivastava, S. 2018. Sentiment classification of online consumer reviews using word vector representations. Procedia Computer Science, 132, 1147–1153.

Cerqueira, A.S., Ferreira, D.D., Ribeiro, M.V., Duque, C.A. 2008. Power quality events recognition using a SVM-based method. Electric Power Systems Research, 78, 1546–1552.

Chang, C.C., Lin, C.J. 2001. LIBSVM: a Library for support vector machines, Software, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

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

Chatterjee, S., Kar, A.K. 2020. Why do small and medium enterprises use social media marketing and what is the impact: Empirical insights from India. International Journal of Information Management, 53, Article 102103.

Chen, J., Huang, H., Tian, S., Qua, Y. 2009. Feature selection for text classification with Naïve Bayes. Expert Systems with Applications, 36, 5432–5435.

Chen, L.-S., Chiu, H.-J. 2009. Developing a neural network based index for sentiment classification. Proceedings of the International MultiConference of Engineers and Computer Scientists, 744–749.

Gokalp, O., Tasci, E., Ugur, A. 2020. A novel wrapper feature selection algorithm based on iterated greedy metaheuristic for sentiment classification. Expert Systems with Applications, 14615, Article 113176.

Hsu, C.-W., Chang, C.-C., Lin, C.-J. 2006. A practical guide to support vector classification. http://www.csie.ntu.edu.tw /~cjlin/ libsvm/index.html.

Karabatak, M., Ince, M.C. 2009. A new feature selection method based on association rules for diagnosis of Erythemato-squamous diseases. Expert Systems with Applications, 36, 12500–12505.

Kennedy, A., Inkpen D. 2006. Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence, 22, 110–125.

Khan, K., Baharudin, B.B., Khan, A., e-Malik, F. 2009. Mining opinion from text documents: A survey, The 3rd IEEE International Conference on Digital Ecosystems and Technologies, 217–222.

Kong, L., Li, C., Ge, J., Zhang, F., Feng, Y., Li, Z., Luo, B. 2020. Leveraging multiple features for document sentiment classification. Information Sciences, 518, 39–55.

Li, B., Xu, S., Zhang, J. 2007. Enhancing clustering blog documents by utilizing author/reader comments. Proceedings of the 45th Annual Southeast Regional Conference, 94–99.

Li, S., Zong, C., Wang, X. 2007. Sentiment classification through combining classifiers with multiple feature Sets. Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering, 135–140.

Liu, B., Hu, M., Cheng, J. 2005. Opinion observer: analyzing and comparing opinions on the web. Proceedings of the 14th international conference on World Wide Web, 342–351.

Martineau, J., Finin, T. 2009. Delta TFIDF: An improved feature space for sentiment analysis. Proceedings of the Third AAAI International Conference on Weblogs and Social Media, San Jose, CA, USA.

Mekawie, N., Hany, A. 2019. Understanding the factors driving consumers' purchase intention of over the counter medications using social media advertising in Egypt: (A Facebook advertising application for cold and Flu products). Procedia Computer Science, 164, 698–705.

Na, J.C., Khoo, C., Wu, P.H.J. 2005. Use of negation phrases in automatic sentiment classification of product reviews. Library Collections, Acquisitions, and Technical Services, 29, 180–191.

O'Keefe, T., Koprinska, I. 2009. Feature selection and weighting methods in sentiment analysis. Proceedings of the 14th Austraasian Document Computing Symposium.

Pang, B., Lee, L., Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. EMNLP, 79–86.

Polat, K., Gunes, S. 2009. A new feature selection method on classification of medical datasets: Kernel F-score feature selection. Expert Systems with Applications, 36, 10367–10373.

Simeon, M., Hilderman, R. 2008. Categorical proportional difference: A feature selection method for text categorization. The Australasian Data Mining Conference, 201–208.

Tan, S., Zhang, J. 2008. An empirical study of sentiment analysis for Chinese documents. Expert Systems with Applications, 34, 2622–2629.

Tang, B., Shepherd, M., Milios, E., Heywood, M.I. 2005. Comparing and combining dimension reduction techniques for efficient text clustering. Proceedings of the Workshop on Feature Selection for Data Mining, SIAM Data Mining.

Tian, X., Tong, W. 2010. An improvement to TF: term distribution based term weight algorithm. The second International Conference on Networks Security Wireless Communications and Trusted Computing, 252–255.

Vapnik, V.N. 1995. The nature of statistical learning theory, Springer-Verlag.

Wang, T., Huang, H., Tian, S., Xu, J. 2010. Feature selection for SVM via optimization of kernel polarization with Gaussian ARD kernels. Expert Systems with Applications, 37, 6663–6668.

Whitelaw C., Garg N., Argamon, S. 2005. Using appraisal groups for sentiment analysis. Proceedings of the 14th ACM international conference on Information and knowledge management, 625–631.

Xu, F., Pan, Z., Xia, R. 2020. E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework. Information Processing & Management, Article 102221.

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

Yu, B., Kaufmann, S., Diermeier, D. 2008. Exploring the characteristics of opinion expressions for political opinion classification. Proceedings of the 9th Annual International Digital Government Research Conference, 82–91.

Zhang, C., Zuo, W., Peng, T., He, F. 2008. Sentiment classification for Chinese reviews using machine learning methods based on string kernel. The Third International Conference on Convergence and Hybrid Information Technology, 909–914.

Zhang, W., Yoshida, T., Tang, X. 2011. A comparative study of TF-IDF, LSI and multi-words for text classification. Expert Systems with Applications, 38, 2758–2765.

Zhang, Y., Zhang, Z., Miao, D., Wang, J. 2019. Three-way enhanced convolutional neural networks for sentence-level sentiment classification. Information Sciences, 477, 55–64.

Zhao, P., Hou, L., Wu, O. 2020. Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification. Knowledge-Based Systems, 1936, Article 105443.