A novel classification approach based on Naïve Bayes for Twitter sentiment analysis

Junseok Song, Kyung Tae Kim, Byungjun Lee, Sangyoung Kim and Hee Yong Youn
College of Software, Sungkyunkwan University
Suwon, Korea
[e-mail: alskpo@skku.edu, kyungtaekim76@gmail.com,{byungjun, impsoft, youn7147}@skku.edu]
*Corresponding author: Hee Yong Youn

Received November 25, 2016; revised February 20, 2017; accepted March 13, 2017; published June 30, 2017

Abstract

With rapid growth of web technology and dissemination of smart devices, social networking service(SNS) is widely used. As a result, huge amount of data are generated from SNS such as Twitter, and sentiment analysis of SNS data is very important for various applications and services. In the existing sentiment analysis based on the Naïve Bayes algorithm, a same number of attributes is usually employed to estimate the weight of each class. Moreover, uncountable and meaningless attributes are included. This results in decreased accuracy of sentiment analysis. In this paper two methods are proposed to resolve these issues, which reflect the difference of the number of positive words and negative words in calculating the weights, and eliminate insignificant words in the feature selection step using Multinomial Naïve Bayes(MNB) algorithm. Performance comparison demonstrates that the proposed scheme significantly increases the accuracy compared to the existing Multivariate Bernoulli Naïve Bayes(BNB) algorithm and MNB scheme.

Keywords: Twitter sentiment analysis, Machine learning, Naive Bayes, Attribute weighting, Feature selection

This research was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No.B0717-17-0070), Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2016R1A6A3A11931385), the second Brain Korea 21 PLUS project. Corresponding author: Hee Yong Youn.

https://doi.org/10.3837/tiis.2017.06.011 **ISSN : 1976-7277**

1. Introduction

Nowadays, social networking service(SNS) is widely used all over the world, and the number of SNS users has been growing dramatically. People use SNS such as Twitter, Facebook, Linked-in, etc. to share their thought, view, and life in online communities, and huge amount of data are created from SNS in real time. Because SNS shows cooperative and interdependent relationship of the individuals of a group, sentiment analysis of SNS [1,2] is an important research for the confirmation of the majority opinion of people and development of intelligent user interface. It is required for providing optimal service and assessing social issue. Critical decision making process also needs to utilize the results of sentiment analysis in various fields. For example, service providers can grasp the response of the users on their services, and manufactures can use it for marketing research. Also, optimal service can be provided to the users with some recommendation system utilizing the sentiment analysis technique [3,4]. Hence, numerous researchers have been attracted to sentiment analysis which involves novel techniques including machine learning [19].

Twitter is one of the most popular SNS platforms used to express opinion, thought, and view of users. The user of Twitter can read and post a 140-character message called a ‘tweet’. The number of monthly active users of Twitter is more than 313 million in 2016 and almost 500 million posted tweets are generated per day [6]. Due to simple and easy accessibility to massive amount of messages generated in real time, Twitter data has been typically adopted as the data set for sentiment analysis. Using various machine learning algorithms, sentiment analysis classifies a Twitter message into ‘positive’ or ‘negative’, and sometimes ‘neutral’.

Machine learning is a powerful technique allowing computer to learn specific topic and predict the result based on that like human [7]. Furthermore, a particular solution can be found using machine learning with exceptional performance compared to human brain. Among various machine learning algorithms, Naïve Bayes algorithm is generally used for the classification problem due to its simplicity and effectiveness [8]. Twitter sentiment analysis with deep learning is also an important issue recently [20]. Deep learning requires relatively large resource in terms of hardware and computation time. This paper targets sentiment analysis with resource constrained system, and thus the scheme based on Naïve Bayes is focused.

A number of approaches have been proposed including attribute weighting, feature selection, and so forth to improve the performance of Naïve Bayes algorithm [9]. Most attribute weighting approaches for text classification utilize a same number of attributes to calculate the weight of each class. In sentiment analysis of Twitter the number of positive words and negative words are different, and the difference can influence the weight of each class. Because the number of words in positive tweets is greater than negative tweets in most cases, the negative words get collateral benefits in attribute weighting. Moreover, the existing feature selection approaches extract a subset of attributes based on the weight of each attribute to enhance the performance of Naïve Bayes. However, these approaches are not suitable for Twitter sentiment analysis since Twitter data has uncountable attributes and contains various meaningless words such as typing error.

In order to overcome the limitation of the existing schemes, new methods of feature selection are proposed for sentiment analysis of Twitter data based on Naïve Bayes algorithm. The first method divides the training set into positive and negative one to calculate the number of

positive words and negative words separately with each set for attribute weighting. The second one utilizes the difference between the weight of positive and negative word for feature selection. Based on the average of the differences of the weights, the weight of some words is changed to zero. This lets meaningless words such as typing error be effectively excluded when the test document is classified. The proposed scheme is evaluated and compared with the existing schemes using 70,000 training document and 3,000 test document obtained from Sentiment140 [10]. The simulation demonstrates that the proposed scheme predicts the class of test set with higher accuracy than the existing approaches. It also identifies that attribute weighting is slightly more influential to the accuracy than feature extraction.

The rest of the paper is organized as follows. Section 2 presents the background and related researches, and Section 3 introduces the proposed scheme. The simulation results of the proposed scheme are presented in Section 4. And finally, Section 5 gives the conclusion and future researches.

2. Related Work

2.1 Twitter Sentiment Analysis

In most sentiment analysis of Twitter, binary classification is generally performed in which the target text is classified as positive, negative, or neutral. In past years a number of researchers have been studying Twitter sentiment analysis using various machine learning techniques. The general structure of Twitter sentiment analysis is shown in **Fig. 1**.

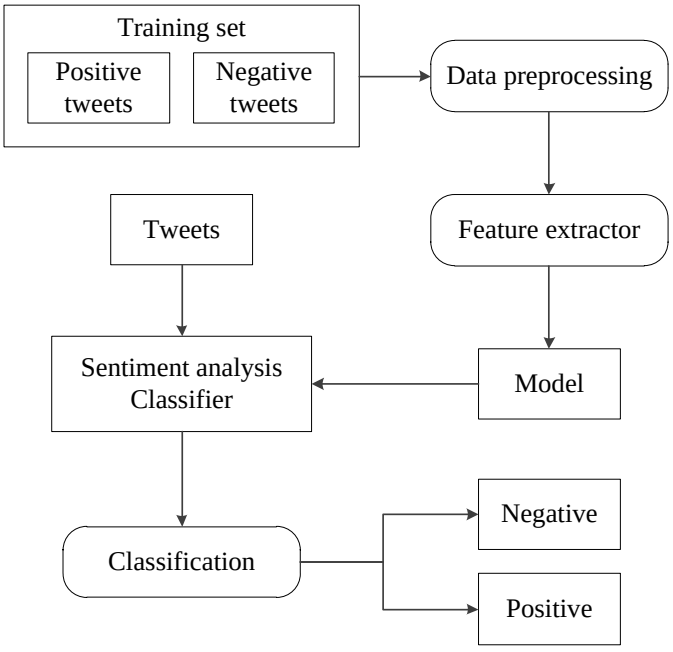


Fig. 1. The structure of Twitter sentiment analysis

An approach for automatic classification of the sentiment of Twitter messages using distant supervision was presented in [6]. The training data of them consists of Twitter message and emoticons used as noisy labels. They used Naïve Bayes, Maximum Entropy(MaxEnt),

Support Vector Machine(SVM) to build models, with 1,600,000 training set and 359 test set extracted from Sentiment140. Their feature extractors are unigrams, bigrams, unigrams and Part Of Speech(POS) tags, and they arrived at a conclusion that POS tags were not useful.

The POS-specific prior polarity features and a tree kernel were presented to remove the necessity for tedious feature engineering and combine many categories of features into one convenient representation [11]. They demonstrated that tree kernel and feature-based approach outperform the unigram baseline using SVM. A feasible solution with good accuracy and time efficiency was also proposed in [5]. Here a new feature combination scheme was developed using the sentiment lexicons and the extracted tweet unigrams of high information gain. Multinomial Naïve Bayes(MNB) was used with 9,000 training set and 1,000 test set, and MNB was found to be the best choice for tweet sentiment analysis. An approach was introduced in [12] which selects a new feature set using information gain, bigram, and object-oriented extraction method. Based on Naïve Bayes and SVM, the accuracy of classifier was shown to be improved through the feature set.

2.2 Naïve Bayes

Naïve Bayes based on Bayes' theorem is a popular algorithm in the classification with a machine learning technique. It is especially useful for learning with high dimensional data like text [13]. Generally, a Bayesian model consists of a structural model and conditional dependencies between random variables [14]. Here an attribute is a meaningful word included in the text to which the classification is made. For example, for the sentence of "I love this brand.", the attributes are "I, love, this, brand". All attributes are assumed to be conditionally independent in Naïve Bayes given the class label. This assumption allows the classifier to be defined as:

$$c(d) = \arg\max_{c \in C} P(c) \prod_{i=1}^{m} P(a_i|c) \tag{1}$$

where d is test document, $c(d)$ is the class of d, and c is class label. The set of all class labels is denoted by C, and m is the number of attributes, and a_i is the value of each attribute, A_i (i = 1,2,…, m). $P(c)$ is the prior probability of class c and $P(a_i|c)$ is the conditional probability of each attribute of the training set.

MNB is a frequency-based model proposed for text classification, where the frequencies of the words in the document are manipulated [15]. Text classification with MNB [16] classifies test document d indicated by a word vector $<w_1,w_2,\ldots, w_m>$ using Eq. (2).

$$c(d) = \arg\max_{c \in C} \left[\log P(c) + \sum_{i=1}^{m} f_i \log P(w_i|c) \right] \tag{2}$$

where m is the number of different words in the document, w_i (i = 1,2,…, m) is the ith word appearing in the document, and f_i (i = 1,2,…, m) is the frequency of w_i in d. In MNB, The prior probability $P(c)$ can be calculated by Eq. (3).

$$P(c) = \frac{\sum_{j=1}^{n} \delta(c_j, c) + 1}{n + l} \tag{3}$$

In addition, the conditional probability $P(w_i|c)$ can be estimated as:

$$P(w_i|c) = \frac{\sum_{j=1}^{n} f_{ji} \delta(c_j, c) + 1}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ji} \delta(c_j, c) + m} \tag{4}$$

where n is the number of training documents, l is the number of classes, f_{ji} is the frequency of w_i in jth training document, c_j is the class of the jth training document. The binary function $\delta(c_j,c)$ can be defined using Eq. (5).

$$\delta(c_j, c) = \begin{cases} 1, & \text{if } c_j = c \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Multivariate Bernoulli Naïve Bayes(BNB) model is another famous statistical language model proposed for text classification. In contrast with MNB, BNB assumes that each feature in a document is described as independent binary variable. BNB considers only the existence of a word in the document without referring to its frequency.

The methods enhancing the performance of Naïve Bayes by machine learning are classified into five categories which are structure extension, feature selection, attribute weighting, local learning, and data expansion [17]. Among these methods, the attribute weighting method of using different weight for each attribute and the feature selection method of selecting a subset of attributes based on the weight are widely employed. Naïve Bayes with attribute weighting can be expressed as:

$$c(d) = \arg\max_{c \in C} P(c) \prod_{i=1}^{m} \mathrm{P}(a_i|c)^{WT_i} \tag{6}$$

where WT_i is the weight of A_i (i = 1,2,…, m). MBN with attribute weighting can be modeled using Eq. (7).

$$c(d) = \arg\max_{c \in C} \left[\log P(c) + \sum_{i=1}^{m} WT_i f_i \log P(w_i|c) \right] \tag{7}$$

where WT_i is the weight of each word w_i (i = 1,2,…, m).

[16] reviewed simple and efficient attribute weighting approaches designed for standard classifiers using Naïve Bayes, and adapted them for Naïve Bayes text classifiers. They used gain ratio-based approach called GRW as one of the attribute weighting approaches for text classification. Their approach assumes that each attribute value is only zero or nonzero. The weight of w_i, WT_i (i = 1,2,…, m), is calculated using a training document set D:

$$WT_i = \frac{IGR(C, w_i) \times m}{\sum_{i=1}^{m} IGR(C, w_i)} \tag{8}$$

where C is the value of the target class, $IGR(C,w_i)$ is the information gain ratio of w_i defined as follows.

$$IGR(C, w_i) = \frac{IG(C, w_i)}{H(w_i)} \tag{9}$$

where $IG(C,w_i)$ indicates the information gain of w_i, $H(w_i)$ is the entropy information of D with regards to w_i. $IG(C,w_i)$ and $H(w_i)$ are computed by Eq. (10) and (11), separately.

$$IG(C, w_i) = H(C) - H(C|w_i) \tag{10}$$

$$H(w_i) = -\sum_{v} \frac{|D_v|}{|D|} \log_2 \frac{|D_v|}{|D|} \tag{11}$$

where $H(C)$ is the entropy of D, $H(C|w_i)$ is the conditional entropy of D given w_i, $|D_v|$ is the size of D, where the number of w_i is v (here $v \in \{0, \bar{0}\}$). The two entropy values can be calculated using Eq. (12) and (13), respectively.

$$H(C) = -\sum_{c} P(c) \log_2 P(c) \tag{12}$$

$$H(C|w_i) = -\sum_{v} \frac{|D_v|}{|D|} \sum_{c} P(c|v) \log_2 P(c|v) \tag{13}$$

Finally, WT_i is used in the analysis not only for the classification of Naïve Bayes text but also conditional probability estimates [18]. Hence, the conditional probability formula of MNB of Eq. (4) is modified as follows.

$$P(w_i|c) = \frac{\sum_{j=1}^{n} WT_i f_{ji} \delta(c_j, c) + 1}{\sum_{i=1}^{m} \sum_{j=1}^{n} WT_i f_{ji} \delta(c_j, c) + m} \tag{14}$$

We next present the proposed scheme which further enhances the accuracy of Twitter sentiment analysis with Naïve Bayes.

3. The Proposed Scheme

In this section the proposed scheme is presented, which exploits a novel attribute weighting and feature selection approach using Naïve Bayes for Twitter sentiment analysis. **Fig. 2** illustrates the operation flow of the proposed approach which adopts two methods based on MNB. The first one aims to calculate the weights more accurately based on the training set divided into positive and negative, while the second one aims to modify the weights using the average of weight differences for automatic feature selection.

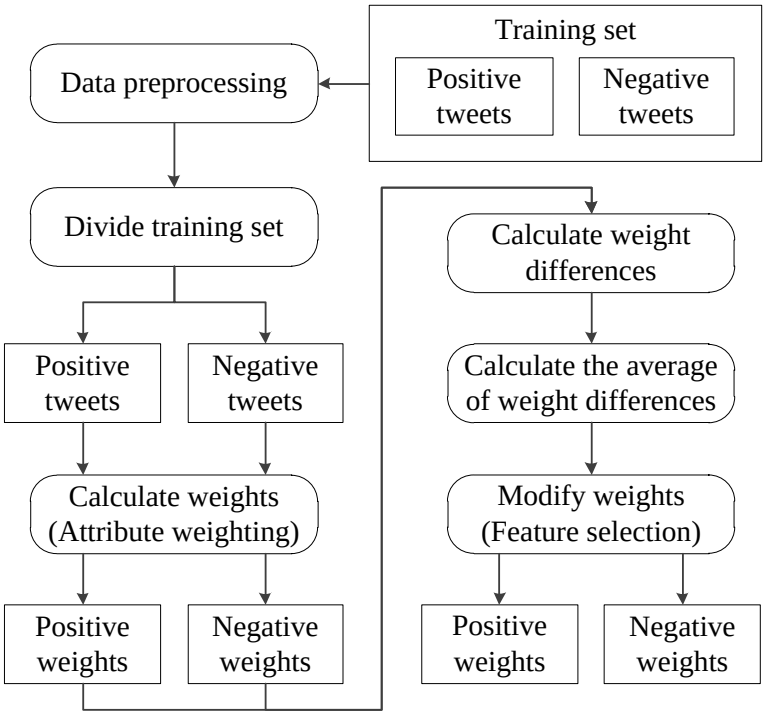


Fig. 2. The operation flow of the proposed approach

3.1 Data Preprocessing

Twitter data contains a lot of words and sentences which are expressed in various ways because of different users [1]. Therefore, unnecessary and redundant data need to be eliminated for sentiment analysis, and the Twitter dataset must be labeled as positive or negative to be applied to the machine learning algorithm. Preprocessing of Twitter data includes the removal of the following objects.

- Non-English tweets
- Numbers
- Hash tags (e.g. #topic), targets (e.g. @username)
- URL and e-mail address
- Special characters including emoticons

3.2 Attribute Weighting

Most existing attribute weighting approaches for text classification utilize a same number of attributes to calculate the weight of each class as Eq. (8) of GRW [16]. In Twitter sentiment analysis the number of positive words and negative words are obviously different, and the difference imposes a significant impact on the weight of each class. In the proposed approach, thus, the training set is first classified into positive and negative, and then the number of positive words and negative words are counted separately from the each of the sets. Hence, a word has only positive (or negative) weight if it exists in only the positive (or negative) set, or both positive and negative weight if it exists in both the set.

The weight of the words in the divided training set, D_c, is calculated using the approach similar to GRW. The weight, $WT_{i,c}$, of each word, $w_{i,c}$ (i = 1,2,…, m_c), in D_c can be defined as:

$$WT_{i,c} = \frac{IGR(c, w_{i,c}) \times m_c}{\sum_{i=1}^{m_c} IGR(c, w_{i,c})} \tag{15}$$

where c is the class label ($\in$ {positive, negative}), m_c is the number of different words in D_c. $IGR(c,w_{i,c})$,which is the information gain ratio of $w_{i,c}$, is obtained by modifying Eq. (9) as follows.

$$IGR(c, w_{i,c}) = \frac{IG(c, w_{i,c})}{H(w_{i,c})} \tag{16}$$

Here $IG(c,w_{i,c})$ and $H(w_{i,c})$ are the information gain and entropy information of $w_{i,c}$, respectively. $IG(c,w_{i,c})$ and $H(w_{i,c})$ are obtained by Eq. (17) and (18).

$$IG(c, w_{i,c}) = H(c) - H(c|w_{i,c}) \tag{17}$$

$$H(w_{i,c}) = -\sum_{v} \frac{|D_{v,c}|}{|D|} \log_2 \frac{|D_{v,c}|}{|D|} \tag{18}$$

where $H(c)$ is the entropy of D_c, $H(c|w_{i,c})$ is the conditional entropy of D_c given the word$w_{i,c}$, $|D_{v,c}|$ is the size of D_c in terms of the number of $w_{i,c}$ for which v ($\in \{0, \bar{0}\}$). $H(c)$ and $H(c|w_{i,c})$ are calculated as:

$$H(c) = -P(c) \log_2 P(c) \tag{19}$$

$$H(c|w_{i,c}) = -\sum_{v} \frac{|D_{v,c}|}{|D|} \sum_{c} P(c|v) \log_2 P(c|v) \tag{20}$$

3.3 Modification of Weights

The existing feature selection approaches use the subset of fixed number of attributes. However, they are not effective for Twitter sentiment analysis since the number of attributes of Twitter data is uncountable. Furthermore, Twitter data contains various meaningless words such as typing error, personal name, and gobbeldy-gook. Typical meaningless words have larger weight value than frequent or important word with the existing attribute weighting approaches because they usually appear infrequently in the training set. The proposed feature selection approach thus uses the difference between the positive weight and negative weight of all the words to modify the weights of meaningless words. After obtaining the differences of the weights of all the words in the training set, the average of the differences is calculated. Based on that, some weight values are changed to zero. This lets the meaningless words be effectively excluded when the class of test document is predicted.

The difference of weight, WD_i, of w_i (i = 1,2,…, m) is defined as follows.

$$WD_i = \begin{cases} \left|WT_{i,p} - WT_{i,n}\right|, & \text{if } w_i \in D_p \text{ and } w_i \in D_n \\ WT_{i,c}, & \text{otherwise} \end{cases} \tag{21}$$

The average of WD_i's, Avg_{WD}, is computed using Eq. (22).

$$Avg_{WD} = \frac{\sum_{i=1}^{m} WD_i}{m} \tag{22}$$

AVG_{WD} and the weight of all words are compared in each training set, D_c. If the weight value is higher than the average, it is modified to zero. Otherwise, it is kept as follows.

$$WT_{i,c} = \begin{cases} 0, & \text{if } WT_{i,c} > Avg_{WD} \\ WT_{i,c}, & \text{otherwise} \end{cases} \tag{23}$$

Finally, the class label of the test document is predicted using non-zero weight words based on Eq. (7). The algorithm of the proposed scheme is presented below.

Algorithm 1. The proposed scheme using GRW (D, d)

Input: a training document set D, a test document d
Output: the class label $c(d)$

Begin
1:Divide D into D_c ($c \in \{\text{positive}, \text{negative}\}$)
2: For $w_{i,c}(i = 1,2, \dots, m_c)$from D_c
Calculate$IGR(c, w_{i,c})$using Eq. (16)
3: For $w_{i,c}(i = 1,2, \dots, m_c)$ from divided D_c
Calculate the weight $WT_{i,c}$ using Eq. (15)
4: For $w_i(i = 1,2, \dots, m)$ from D
Calculate the weight difference, WD_i, using Eq. (21)
5: Calculate the average of WD_i's using Eq. (22)
6: For $w_{i,c}(i = 1,2, \dots, m_c)$ from D_c
Modify $WT_{i,c}$ using Eq. (23)
7: For d
(a) Calculate $P(c)$using Eq. (3)
(b) Calculate $P(w_i|c)$ using Eq. (14)
(c) Predict $c(d)$ using Eq. (7)
8:Return $c(d)$
End

3.4 Case Study

Here the computation steps of the existing MNB scheme and the proposed scheme are compared with a case study. **Table 1** is the data used for the case study. With the existing MNB scheme, the frequency of each word (not, do, forget, ever, cheer, never) in the training document set, D, is computed. Six documents are in D (3 positive ones and 3 negative ones), and the class label of test document, d_1, is positive.

Table 1. The data used for case study

Word	**Training document**					
	D_1	D_2	D_3	D_4	D_5	D_6
Not	0	0	0	2	0	1
Do	1	2	1	1	1	1
Forget	0	1	0	1	1	1
Ever	1	1	2	1	2	0
Cheer	0	1	2	0	0	0
Never	0	0	0	1	0	2
Class	Positive			Negative		
Test document(d_1)	Do not ever forget you are never alone cheer up					

The class label of d_1 is predicted with MNB using Eq. (2). $log(P(pos|d_1))$ and $log(P(neg|d_1))$ are calculated as:

$$\begin{aligned} logP(pos \mid d_1) &= logP(not \mid pos) + logP(do \mid pos) \\ &\quad + logP(forget \mid pos) + logP(ever \mid pos) \\ &\quad + logP(cheer \mid pos) + logP(never \mid pos) \\ &\quad + logP(pos) \\ &= -5.53164 \end{aligned} \tag{24}$$

$$\begin{aligned} logP(neg \mid d_1) &= logP(not \mid neg) + logP(do \mid neg) \\ &\quad + logP(forget \mid neg) + logP(ever \mid neg) \\ &\quad + logP(cheer \mid neg) + logP(never \mid neg) \\ &\quad + logP(neg) \\ &= -5.22405 \end{aligned} \tag{25}$$

Since $log(P(pos|d_1)) < log(P(neg|d_1))$, the class label of d_1 is predicted to be negative. Note that the nuance of the test document is not negative but positive, and the prediction is wrong. We next show how it is predicted using the proposed scheme.

For the same input data, $WT_{i,c}$ of each word, $w_{i,c}$ ($i = 1,2,\ldots, m_c$), in divided training set D_c and the weight difference, WD_i, are computed using Eq. (15) and Eq. (21) with the proposed scheme. **Table 2** shows the weights and the weight differences of the words in D.

Table 2. The weights with the proposed scheme

Word	**Weight**		**Weight difference**
	Positive	**Negative**	
Not	0	1.531016	1.531016
Do	0.840682	0.15758	0.683102
Forget	1.738987	1.023733	0.715254
ever	0.129831	0.756655	0.626824
cheer	1.290499	0	1.290499
never	0	1.531016	1.531016

The average of weight differences, Avg_{WD}, is calculated using Eq. (22).

$$Avg_{WD} = 1.062952 \tag{26}$$

Based on Avg_{WD}, $WT_{i,c}$ is modified. Table 3 shows the weights of the words in D after the modification of the weights.

Table 3. The weights after modification

Word	Weight	
	Positive	**Negative**
not	0	0
do	0.840682	0.15758
forget	0	1.023733
ever	0.129831	0.756655
cheer	0	0
never	0	0

Finally, $log(P(pos|d_1))$ and $log(P(neg|d_1))$ are estimated as:

$$\begin{aligned} logP(pos \mid d_1) &= WT_{do,pos} \times logP(do \mid pos) \\ &+ WT_{ever,pos} \times logP(ever \mid pos) \\ &+ logP(pos) \\ &= -0.89464 \end{aligned} \tag{27}$$

$$\begin{aligned} logP(neg \mid d_1) &= WT_{do,pos} \times logP(do \mid neg) \\ &+ WT_{forget,pos} \times logP(forget \mid neg) \\ &+ WT_{ever,pos} \times logP(ever \mid neg) \\ &+ logP(neg) \\ &= -1.66347 \end{aligned} \tag{28}$$

Since $log(P(pos|d_1)) > log(P(neg|d_1))$, the class label of d_1 is predicted to be positive with the proposed scheme. The effectiveness of the proposed scheme is evaluated next by computer simulation.

4. Performance Evaluation

In this section the accuracy of Twitter sentiment analysis using the proposed scheme is examined through computer simulation. It is also compared with the existing MNB and BNB scheme. The accuracy of Twitter sentiment analysis is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{29}$$

Here TP, TN, FP, FN are the number of true positive, true negative, false positive, and false negative documents, respectively. **Table 4** shows the general confusion matrix.

Table 4. Confusion matrix

Actual \ Predicted	**Positive**	**Negative**
Positive	TP	FN
Negative	FP	TN

To evaluate the accuracies, the subsets of 1,600,000 data set supplied by Sentiment140 are used in the computer simulation. The total data set contains 800,000 tweets of positive class and 800,000 tweets of negative class. In this experiment the ratio of positive and negative classes is 5:5 in training and test sets.

Table 5 is the number of words in the training sets. The size of each training set is 10,000, 20,000, 30,000, 40,000, 50,000, and 70,000. After the preprocessing of the training set, the number of words in the training sets is dropped about 38~53%. Also, the decrease in the number of positive words is about 2~4% larger than negative words.

Table 5. The number of words in the training sets

The number of documents		Before preprocessing	After preprocessing
10,000	Total	25,242	14,438
	Positive	15,532	9,357
	Negative	13,939	8,698
20,000	Total	42,747	22,763
	Positive	26,495	13,793
	Negative	23,380	14,818
30,000	Total	58,145	29,728
	Positive	36,008	19,250
	Negative	31,758	18,066
40,000	Total	72,049	35,649
	Positive	44,707	23,053
	Negative	39,132	21,642
50,000	Total	85,138	40,999
	Positive	52,740	26,412
	Negative	46,223	25,031
70,000	Total	110,384	51,141
	Positive	68,814	33,053
	Negative	59,394	31,026

Fig. 3 shows the accuracies of Twitter sentiment analysis with the proposed scheme using the 1,000 test set and different training sets. The accuracies with the MNB, BNB scheme are also compared. Observe from the figure that the accuracy of the proposed scheme is higher than the other schemes. As the size of training set increases, the accuracy of all the schemes also increases as expected.

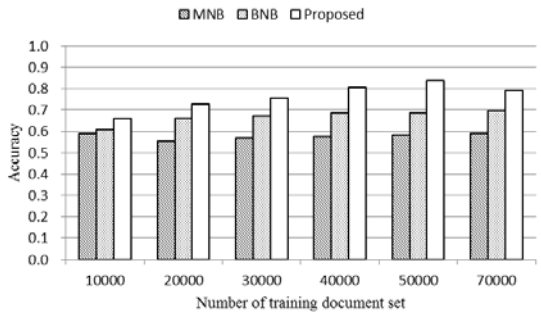


Fig. 3. The accuracies of Twitter sentiment analysis for 1,000 test set

Fig. 4 and **Fig. 5** show the accuracies with 2,000 test set and 3,000 test set, respectively. Notice from the figures that similar results as with the 1,000 test set were achieved.

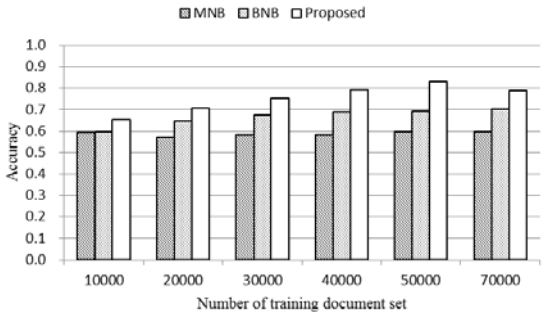


Fig. 4. The accuracies of Twitter sentiment analysis for 2,000 test set

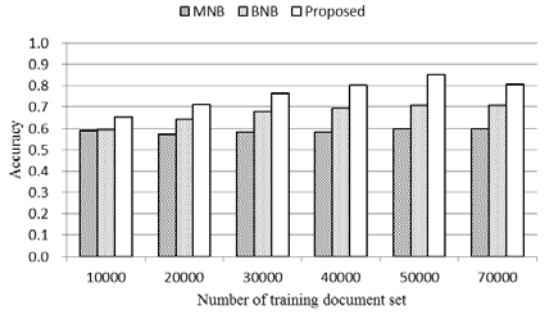


Fig. 5. The accuracies of Twitter sentiment analysis for 3,000 test set

As mentioned earlier, attribute weighting and feature selection are important steps in sentiment analysis, and thus we have proposed a new approach for each of them for improving the accuracy. **Fig. 6** shows relative effectiveness of them with 1,000 test set. Here the bars of attribute weighting are for the case of sentiment analysis based on the proposed attribute weighting approach without feature selection. The bars of feature selection are with the existing attribute weighting approach and the proposed feature selection approach. The data with both of the proposed approached are shown with white bars. Notice from the figure that feature selection is always more influential to the accuracy than attribute weighting regardless of the number of training documents. Also, the proposed attribute weighting approach is always more effective than the existing approach as identified by comparing the bars of 'Feature selection' and 'Proposed'.

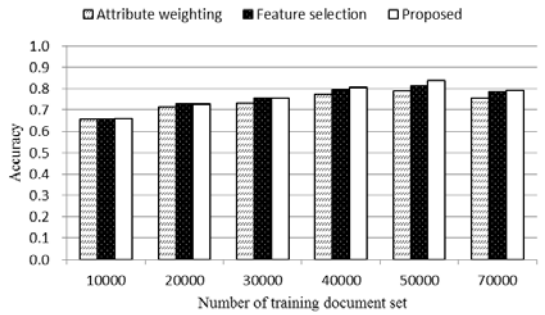


Fig. 6. The accuracies of the proposed approaches for 1,000 test set

Fig. 7 and **Fig. 8** are the accuracies comparing the attribute weighting approach and feature selection approach using 2,000 test set and 3,000 test set, respectively. Notice from the figures that similar results as case of the 1,000 test set were achieved.

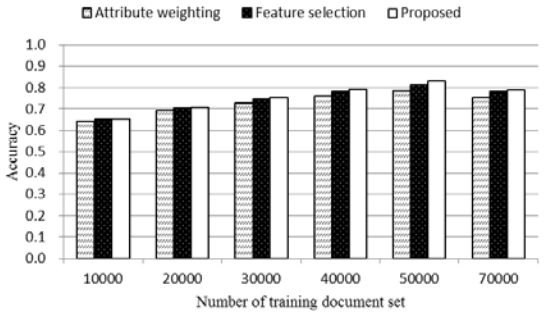


Fig. 7. The accuracies of the proposed approaches for 2,000 test set

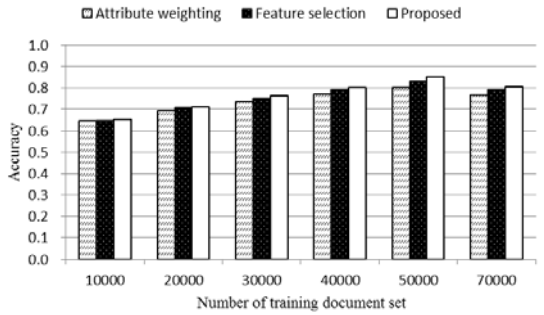


Fig. 8. The accuracies of the proposed approaches for 3,000 test set

Fig. 9 investigates the performances of the schemes for the positive sets and negative sets separately using 50,000 training set and 3,000 test set. Observe that the proposed approaches consistently excel the existing MNB and BNB scheme. Also notice that the accuracies for positive sets are higher than for negative sets. This is because the document is deemed to be positive if the basis for classifying the test document is insufficient.

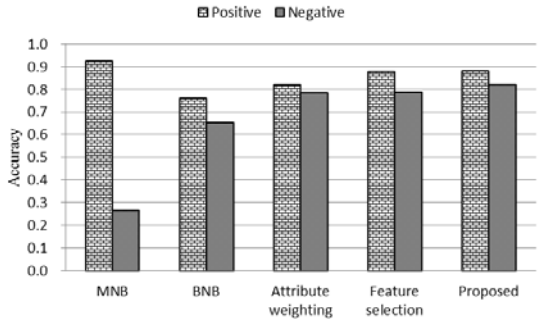


Fig. 9. The accuracies for positive and negative sets separately

Table 6 compares the proposed scheme with the existing schemes in various aspects. Here the maximum achievable accuracies of the previous schemes were quoted from the papers. Notice that the experimental settings of them are different. Therefore, for fair comparison, the

maximum achievable accuracies of each of the schemes are used.

Table 6. Comparison of the proposed scheme and the previous schemes

		[6]	[11]	[12]	[5]	Ours
Maximun accuracy		83.00%	75.39%	80.00%	84.60%	85.33%
Classifier		MaxEnt	SVM	MNB	MNB	MNB
Size	**Training set**	1,600,000	5,127	20,000	9,000	50,000
	Test set	359		500	1,000	3,000
Sentiment(Neutral)		O	X	O	X	X

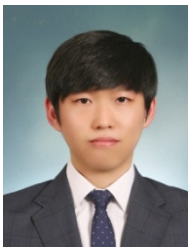
5. Conclusion

In this paper a novel attribute weighting and feature selection approach for Twitter sentiment analysis have been presented based on Naïve Bayes. Most attribute weighting and feature selection approaches based on the Naïve Bayes algorithm employ a same number of attributes to estimate the weight of each class. As a result, if the number of attributes of each class is different, the difference can influence the weight of each class. Since the number of words in positive tweets is greater than negative tweets in most cases, the negative words get collateral benefits in attribute weighting. Moreover, the existing feature selection approaches are not effective to consider uncountable attributes and meaningless attributes. These result in decreased accuracy of Naïve Bayes in Twitter sentiment analysis. Two methods were proposed to resolve these issues. The first method effectively reflects the difference in the number of positive words and the number of negative words in calculating the weights, while the second one identifies significant words to predict the class of test document. According to the experiment with actual test documents, the proposed approach consistently allows higher accuracy than the existing Naïve Bayes based approaches for Twitter sentiment analysis.
As future study the proposed scheme will be enhanced by applying various N-gram problems such as bigram and trigram using more sophisticated attribute weighting and feature selection approach. The effectiveness of the proposed approach will also be investigated with the application to other classification problem such as text classification and traffic classification.

References

[1] Sitaram Asur and Bernardo A. Huberman, “Predicting the Future with Social Media,” in *Proc. of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pp.492-499, 2010. Article (CrossRef Link)
[2] Jeffrey Nichols, Jalal Mahmud and Clemens Drews, “Summarizing Sporting Events Using Twitter,” in *Proc. of the 2012 ACM international conference on Intelligent User Interfaces*, pp.189-198, 2012. Article (CrossRef Link)
[3] Anurag P. Jain and Vijay D. Katkar, “Sentiments analysis of Twitter data using data mining,” in *Proc. of International Conference on Information Processing*,pp.807-810, 2015. Article (CrossRef Link)
[4] Vishal A. Kharde and S.S. Sonawane, “Sentiment Analysis of Twitter Data: A Survey of Techniques,” *International Journal of Computer Applications*, vol. 139, no. 11, pp.5-15, April 2016. Article (CrossRef Link)
[5] Ang Yang, Jun Zhang, Lei Pan and Yang Xiang, “Enhanced Twitter Sentiment Analysis by Using Feature Selection and Combination,” in *Proc. of International Symposium on Security and Privacy in Social Networks and Big Data*, pp.52-57, 2015. Article (CrossRef Link)
[6] Alec Go, Richa Bhayani andLei Huang, “Twitter Sentiment Classification using Distant Supervision,”*CS224N Project Report, Stanford*. 1, 2009. Article (CrossRef Link)

[7] Fabrizio Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Survey*, vol. 34, no. 1, pp.1-47, March, 2002. Article (CrossRef Link)
[8] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Informatica*, vol. 31, no. 3, pp.249-268, 2007. Article (CrossRef Link)
[9] Jingnian Chen, Houkuan Huang, Shengfeng Tian and Youli Qu, "Feature selection for text classification with Naïve Bayes," *Expert Systems with Applications*, vol. 36, no. 3, pp.5432-5435, April, 2009. Article (CrossRef Link)
[10] Saif M. Mohammad, Svetlana Kiritchenko and Xiaodan Zhu, "NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets," in *Proc. of the seventh international workshop on Semantic Evaluation Exercises*, 2013. Article (CrossRef Link)
[11] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau, "Sentiment analysis of Twitter data," in *Proc. of the Workshop on Languages in Social Media*, pp.30-38, 2011. Article (CrossRef Link)
[12] Bac Le and Huy Nguyen, "Twitter Sentiment Analysis Using Machine Learning Techniques," A*dvanced Computational Methods for Knowledge Engineering*, pp.279-289, 2015. Article (CrossRef Link)
[13] Jia Wu, Shirui Pan, Xingquan Zhu, Zhihua Cai, Peng Zhang and Chengqi Zhang, "Self-adaptive attribute weighting for Naive Bayes classification," *Expert Systems with Applications*, vol. 42, no. 3, pp.1487-1502, February, 2015. Article (CrossRef Link)
[14] Nir Friedman, Dan Geiger and Moises Goldszmidt, "Bayesian Network Classifiers," *Machine Learning*, vol. 29, no. 2, pp.131-163, November, 1997. Article (CrossRef Link)
[15] Andrew McCallum and Kamal Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," in *Proc. of AAAI-98 workshop on learning for text categorization*, pp. 41-49, 1998. Article (CrossRef Link)
[16] Lungan Zhang, Liangxiao Jiang, Chaoqun Li and Ganggang Kong, "Two feature weighting approaches for naive Bayes text classifiers," *Knowledge-Based Systems*, vol. 100, no. 15, pp.137-144, May, 2016. Article (CrossRef Link)
[17] Liangxiao Jiang, Harry Zhang andZhihua Cai, "A Novel Bayes Model: Hidden Naive Bayes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 10, pp.1361-1371, October, 2009. Article (CrossRef Link)
[18] Liangxiao Jiang, Chaoqun Li, Shasha Wang and Lungan Zhang, "Deep feature weighting for naive Bayes and its application to text classification," *Engineering Applications of Artificial Intelligence*, vol. 52, pp.26-39, June, 2016. Article (CrossRef Link)
[19] Xuemeng Song, Zhao-Yan Ming, Liqiang Nie, Yi-Liang Zhao and Tat-Seng Chua, "Volunteerism Tendency Prediction via Harvesting Multiple Social Networks," *ACM Transactions on Information Systems*, vol. 34, no. 2, pp.1-27, April, 2016. Article (CrossRef Link)
[20] Aliaksei Severyn and Alessandro Moschitti, "Twitter Sentiment Analysis with Deep Convolutional Neural Networks," in *Proc. of International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 959-962, August, 2015. Article (CrossRef Link)



Junseok Song received the BS degree in computer science from Seokyeong University, Seoul, Korea, in 2015. He is currently working toward the MS degree in College of Software in Sungkyunkwan University and research staff of Ubiquitous computing Technology Research Institute. His current research interests include Internet of Things Technology, Big data, and Machine learning.

Kyung Tae Kim received the PhD degree in College of Information and Communication Engineering from Sungkyunkwan University, Korea, in 2013. He is currently a research professor at the college of software from Sungkyunkwan University, Korea. His current research interests include Ubiquitous computing, Wireless Networks, and Internet of Things Technology.

Byungjun Lee received the MS degree in College of Information and Communication Engineering from Sungkyunkwan University, Korea, in 2012. He is currently working toward the PhD degree in College of Software in Sungkyunkwan University and research staff of Ubiquitous computing Technology Research Institute. His current research interests include Internet of Things Technology, Wireless Networks, and Machine learning.

Sangyoung Kim received the BS degree in computer engineering from Inje University, Gyeongsangnam-do, Korea, in 2015. He is currently working toward the MS degree in College of Software in Sungkyunkwan University and research staff of Ubiquitous computing Technology Research Institute. His current research interests include Internet of Things Technology, Big data, and Software Defined Networking.

Hee Yong Youn received the BS and MS degree in electrical engineering from Seoul National University, Seoul, Korea, in 1977 and 1979, respectively, and the PhD degree in computer engineering from the University of Massachusettsat Amherst, in 1988. He had been Associate Professor of Department of Computer Science and Engineering, The University of Texas at Arlington until 1999. He is presently Professor of College of Software, Sungkyunkwan University, Suwon, Korea, and Director of Ubiquitous computing Technology Research Institute. He has been also Consulting Professor of Software R&D Center, Device Solutions, Samsung Electronics, Korea. His research interests include distributed and ubiquitous computing, IoT, and intelligent system. He has published more than 400 papers in int'l journals and conference proceedings, and received Outstanding Paper Award from the 1988 IEEE International Conference on Distributed Computing Systems, 1992 Supercomputing, and 2012 IEEE Int'l Conf. on Computer, Information and Telecommunication Systems, 2014 The 6th International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, respectively. Dr. Youn has also been General Chair of IEEE PRDC 2001, Int'l Conf. on Ubiquitous Computing Systems (UCS) in 2006 and 2009, UbiComp 2008, CyberC 2010, Program Chair of PDCS 2003 and UCS 2007. Dr. Youn is a senior member of the IEEE Computer Society.