IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 25, 2016, accepted January 20, 2017, date of publication February 22, 2017, date of current version March 28, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2672677

Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis

ZHAO JIANQIANG[1,2,3] and GUI XIAOLIN[1,3]
[1]School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[2]Xi'an Politics Institute, Xi'an, Shaanxi Province, 710068, China
[3]Key laboratory of Computer Network of Shaanxi Province, Xi'an 710049, China

Corresponding author: G. Xiaolin (xlgui@mail.xjtu.edu.cn)

This work was supported in part by the NSFC under Grant 1472316, in part by the Shaanxi Science and Technology Plan Project under Grants 2016ZDJC-05 and 2013SZS16-Z01/P01/K01, and in part by the Fundamental Research Funds for Ministry of Education of China under Grant XKJC2014008.

ABSTRACT Twitter sentiment analysis offers organizations ability to monitor public feeling towards the products and events related to them in real time. The first step of the sentiment analysis is the text pre-processing of Twitter data. Most existing researches about Twitter sentiment analysis are focused on the extraction of new sentiment features. However, to select the pre-processing method is ignored. This paper discussed the effects of text pre-processing method on sentiment classification performance in two types of classification tasks, and summed up the classification performances of six pre-processing methods using two feature models and four classifiers on five Twitter datasets. The experiments show that the accuracy and F1-measure of Twitter sentiment classification classifier are improved when using the pre-processing methods of expanding acronyms and replacing negation, but barely changes when removing URLs, removing numbers or stop words. The Naive Bayes and Random Forest classifiers are more sensitive than Logistic Regression and support vector machine classifiers when various pre-processing methods were applied.

INDEX TERMS Twitter, sentiment analysis, text pre-processing.

I. INTRODUCTION

Twitter, with over 313 million[1] monthly active users and over 500 million tweets per day,[2] has now become a goldmine for organizations and individuals who have a strong social, political or economic interest in maintaining and enhancing their clout and reputation. Sentiment analysis offers these organizations the ability to monitor different social media sites in real time.

Sentiment analysis is the process of automatically detecting whether a text segment contains emotional or opinionated content, and it can furthermore determine the text's polarity. Twitter sentiment classification aims to classify the sentiment polarity of a tweet as positive, negative or neutral.

Tweets are usually composed of incomplete, noisy and poorly structured sentences, irregular expressions, ill-formed words and non-dictionary terms. Before feature selection, a series of pre-processing (e.g., removing stop words, removing URLs, replacing negations) are applied to reduce the amount of noise in the tweets. Pre-processing is performed extensively in existing approaches, especially in machine learning-based approaches [1]–[4]. However, few studies focus on the effect of pre-processing method on the performance of Twitter sentiment analysis. This paper concentrates on exploring various pre-processing methods for elevating the performance of Twitter sentiment analysis.

This paper evaluated the effects of various pre-processing methods on sentiment classification, including removing URLs, replacing negation, reverting repeated letters, removing stop words, removing numbers and expanding acronyms. We used two feature models and four classifiers to identifying tweet sentiment polarity on five Twitter datasets. The experimental results show that the performance of sentiment classification improves after expanding acronyms and replacing negation, but barely changes when removing URLs, removing stop words or numbers.

The remainder of this paper is structured as follows. Related studies and background are discussed in Section II. The evaluation approach is presented in Section III. The experimental results are presented in Section IV. Discussion is covered in Section V. Finally, in Section VI, we conclude the paper.

[1]http://en.wikipedia.org/wiki/Twitter
[2]http://blog.Twitter.com/2014/the-2014-yearontwitter

II. RELATED WORK AND BACKGROUND

Most existing approaches [1]–[8] to identify the sentiment polarity of tweets apply text pre-processing (e.g., POS, removing URLs, expanding acronyms, replacing negative mentions, stemming, removing stop words) to reduce the amount of noise in the tweets. The hypothesis is that data pre-processing reduces the noise in the text, and it should help to improve the performance of the classifier and speed up the classification process.

Haddi *et al*. [9] explored the role of text pre-processing in movie reviews sentiment analysis. The experimental results show that the accuracy of sentiment classification may be significantly improved using appropriate features and representation after pre-processing. Saif *et al*. [4] studied the effect of different stop words removal methods for polarity classification of tweets and whether removing stop words affects the performance of Twitter sentiment classifiers. They applied six different stop words identification methods to six different Twitter datasets and observed how removing stop words affects two supervised sentiment classification methods. They assessed the impact of removing stop words by observing fluctuations on the level of data sparsity, the size of the classifier's feature space and its classification performance. Using pre-compiled lists of stop words negatively impacted the performance of Twitter sentiment classification approaches. Saif *et al*. [5] found that pre-processing led to a significant reduction of the original feature space. After pre-processing, the vocabulary size was reduced by 62%. However, they did not discuss the effect on the performance of Twitter sentiment classifiers. Bao *et al.* [10] explored the effect of pre-processing methods on Twitter sentiment classification. They evaluated the effects of URLs, negation, repeated letters, stemming and lemmatization. Experimental results on the Stanford Twitter Sentiment Dataset show that sentiment classification accuracy increases when URL features reservation, negation transformation and repeated letters normalization are employed, but decreases when stemming and lemmatization are applied. Zhao [11] evaluated the accuracy of URLs, stop words, repeated letters, negation, acronym and numbers in the binary Twitter sentiment classification task. The experiments show that the accuracy of sentiment classification rises after expanding acronym and replacing negation, although hardly change when removal URL, removal numbers and removal stop words are applied.

From the above reviews, there is a lack of proper and deep analysis of the impact of text pre-processing on Twitter sentiment classification. To fill this gap, this paper focus on evaluating the effects of text pre-processing on Twitter sentiment classification using different feature models and machine learning classifiers on five Twitter datasets in two types of classification task.

III. PRE-PROCESSING ANALYSIS SETUP

To assess the effect of various pre-processing method, six pre-processing methods are applied to sentiment classification using four different classifiers on five Twitter datasets. The complete analysis setup is composed as follows.

A. PRE-PROCESSING

The pre-processing methods that are assessed in this paper are as follows:

— Replacing negative mentions. Tweets consist of various notions of negation. In general, negation plays an important role in determining the sentiment of the tweet. Here, the process of negation is transforming ''won't'', ''can't'', and ''n't'' into ''will not'', ''cannot'', and ''not'', respectively.

— Removing URL links in the corpus. Most researchers consider that URLs do not carry much information regarding the sentiment of the tweet. Here, Twitter's short URLs are expanded to URLs and are tokenized. Then, the URL matching the tokens are removed from tweets to refine the tweet content.

— Reverting words that contain repeated letters to their original English form. Words with repeated letters, e.g., ''coooool'', are common in tweets, and people tend to use this way to express their sentiments. Here, a sequence of more than three similar characters is replaced by three characters. For example, ''cooooool'' is replaced by ''coool''. Using three characters distinguish words like ''cool'' from ''cooooool''.

— Removing numbers. In general, numbers are of no use when measuring sentiment and are removed from tweets to refine the tweet content.

— Removing stop words. Stop words usually refer to the most common words in a language, such as ''the'', ''is'', and ''at''. Most researchers consider that stop words play a negative role in the task of sentiment classification, and they are removed before feature selection by researchers. The classic method of removing stop words is the method based on pre-compiled lists. Multiple lists exist in the literature [12], [13]. The classic Van stoplist [12] is selected in this paper.

— Expanding acronyms to their original words by using an acronym dictionary. Acronyms and slang are common in tweets but are ill-formed words. It is necessary to expand them to their original words. This paper expands the acronyms and slang to their original words using the acronym dictionary Internet Slang Dict.[3] Internet Slang Dict consists of slang and acronyms that users have created as an effort to save keystrokes. Terms have originated from various sources, including Bulletin Boards, AIM, Yahoo, IRC, Chat Rooms, Email, and Cell Phone Text Messaging. Each acronym corresponds to an explanation. Example, ''*4 u'' is ''Kiss for you'', ''2 mro'' is ''tomorrow''.

B. FEATURE MODELS

1) WORD N-GRAMS FEATURES MODEL

Word n-grams features are the simplest feature for Twitter sentiment analysis. Researchers report state-of-the-art performance for sentiment analysis on Twitter data using a unigram

[3]Acronym list. [Online]. Available: http://www.noslang.com/dictionary/

model [14], [15]. In this work, word unigram and bigram feature (referred to N-grams model) is one of the feature models.

2) PRIOR POLARITY SCORE FEATURE MODEL

A prior polarity score is a lexicon-based sentiment feature, and some approaches [16]–[18] commonly use it as a sentiment feature for tweet sentiment analysis. We used the AFINN[4] lexicon and extended it using Senti-Wordnet [19] to obtain the prior polarity score of the tweet. The prior polarity score of a tweet is the sum of the sentiment score of each word in the tweet. The sentiment score of each word is computed by measuring the *PMI* (point-wise mutual information) between the term and the positive or negative category of the tweet using the formula:

$$SenScore(\mathrm{w}) = PMI(w, pos) - PMI(w, neg)$$

Where w is a term in the lexicon, $PMI(w, pos)$ is the *PMI* score between w and the positive class, and $PMI(w, neg)$ is the *PMI* score between w and the negative class. Therefore, a positive $SenScore\ (w)$ suggests a stronger association of word w with positive sentiment and vice versa.

In this work, the prior polarity score feature (referred to Prior polarity model) is another feature model.

C. TWITTER SENTIMENT CLASSIFIERS

To assess the effect of pre-processing on sentiment classification, we used four popular supervised classifiers in the literature of sentiment analysis, Support Vector Machine (SVM, parameter c is 100, gamma is 0.5, kernel is linear, other parameters are the default values), Naive Bayes (NB), Logistic Regression (LR, default parameters), and Random Forest (RF, parameter max_depth is 30, n_estimators is 4000, other parameters are set to the default values). This paper uses the GridSearch search for these parameters as the optimal parameters and uses scikit-learn library to perform the classifier.

D. BASELINE AND EVALUATION CRITERIA

Two classification tasks are performed: a binary task of classifying sentiment into positive and negative classes and a 3-way task of classifying sentiment into positive, negative, and neutral classes. The binary task is performed on all five datasets and the 3-way task is performed on four datasets using SVM, NB, LR, and RF classifiers. The baseline method is the classic method (C-Method) respectively using the N-grams and the Prior polarity model, which was applied all six pre-processing methods, including removing URLs, removing stop words, removing numbers, reverting words that contain repeated letters to their original form, replacing negative mentions, and expanding acronyms to original word.

The accuracy and F1-measure are used to evaluate the overall sentiment classification performance. The effect of text pre-processing is evaluated by the gain or loss of accuracy and F1- measure. In the binary task, the F1-measure is the average of positive and negative classification F1-measure. In the 3-way task, the F1-measure is average of the positive, neutral, and negative classification F1-measure. We use 10-fold cross validation, which is a technique that is useful to evaluate a classification algorithm for a given corpora, splitting and evaluating the training set several times.

E. DATASETS

Pre-processing may have different impacts in various contexts. Words and URLs that do not provide any discriminative power in one context may carry some semantic information in another context. This paper studies the effects of pre-processing on five different Twitter datasets that have been used in other sentiment analysis literature.

The Stanford Twitter Sentiment Test (STS-Test) dataset was introduced by Go *et al.* [14]. It has been manually annotated and contains 177 negative, 182 positive and 139 neutrals tweets. Although the Stanford test set is relatively small, it has been widely used in the literature [5], [20] for different evaluation tasks.

SemEval2014 dataset was provided in SemEval2014 Task9.[5] The dataset consists of tweet id's which have been annotated with positive, negative and neutral labels. Some of the tweets were not available for downloading. This leaves us with 11042 tweets for testing.

The Stanford Twitter Sentiment Gold (STS-Gold) dataset was introduced by Saif *et al.* [21]. The dataset has been manually annotated both the tweet-level and the entity-level by three graduate students.

The Sentiment Strength Twitter Dataset (SS-Twitter) consists of 4242 tweets manually labeled with their positive and negative sentiment strengths. The dataset was constructed by Thelwall *et al.* [17] to evaluate SentiStrenth.

The Sentiment Evaluation Dataset (SE-Twitter) was introduced by Sacha Narr *et al.* [22]. The dataset consists of 6745 tweets that have been human-annotated with sentiment labels by three Mechanical Turk workers.

Table 1 displays the distribution of tweets in the five selected datasets according to these sentiment labels.

TABLE 1. Total number of tweets and the tweet sentiment distribution in all datasets.

Dataset	No.of Tweets	#Negative	#Neutral	#Positive
STS-test	498	177	139	182
SemEval2014	11042	1650	5150	4242
STS-Gold	2034	1402	--	632
SS-Twitter	4242	1037	1953	1252
SE-Twitter	6745	990	4097	1658

[4]Afinn-111. [Online]. Available: http://www2.imm.dtu.dk/pubdb/views/publication details.php?id=6010

[5]Dataset.Online Available: http://alt.qcri.org/semeval2014/task9/index.php?id=data-and-tools

IV. EXPERIMENTAL RESULTS

This section reports the results obtained before and after several types of pre-processing methods are applied individually. The baseline (C-Method) is the method that was applied all six pre-processing methods. The impact of pre-processing is assessed by observing fluctuations (increases and decreases) on classification performance of the sentiment classification task, which is measured in terms of accuracy and average F1-measure. The accuracy improvement of one pre-processing method was calculated as:

$$\text{Accuracy}_{\text{improvement}} = \text{Accuracy}_{\text{baseline}} - \text{Accuracy}_{\text{compared}}$$

The average F1-measure improvement of one pre-processing method was calculated as:

$$\text{F1}_{\text{improvement}} = \text{Average F1}_{\text{baseline}} - \text{Average F1}_{\text{compared}}$$

For example, for removing URLs method, the compared method is the method that was applied other five pre-processing except for removal of URLs.

Table 2 reports the effect of the removal of URLs on classification performance and is the result of a comparison of the performance of the baseline and the method that was applied other five pre-processing method except for removal of URLs. It can be observed from Table 2 that the performance of every classifier in the N-grams model does not change after removing the URLs. In the Prior polarity model, removing URLs slightly reduces the accuracy and average F1-measure of SVM on the STS-Test datasets and slightly improves on the SemEval2014 dataset, and fluctuation range is limited to 0.3%. In the binary classification task, a small loss in accuracy and average F1-measure is encountered when using NB in the Prior polarity model on the SemEval2014 dataset. In the binary classification task, the accuracy and average F1-measure of LR do not change on all datasets in the Prior polarity model. In the 3-way classification task the performance of LR fluctuates from −0.18% to 0.11%. The performance of RF fluctuates before and after removing URL, and the range is from −0.50% to 0.29% in the Prior polarity model on all datasets. One factor of the fluctuating of RF performance is the initial random value of the classifiers.

Table 2 indicates that URLs do not contain useful information for sentiment classification. This conclusion is not consistent with the experiment results of Bao *et al*. [10]. Their experiments results on Stanford Twitter Sentiment Dataset show that URLs features reservation have a positive impact on classification accuracy.

To estimate true value of removal of URLs, removal of numbers, removal of stopwords, we check the effect of intentionally neutral modifications. We used random deletion of one word as neutral modifications. For assessment of the effect of removal URLs,we use random deletion of one word to replace the removal of URLs to evaluate the effect of the removal of URLs. Table 3 is the result of a comparison of the performance of the baseline and the removal URLs replaced by random deletion of one word in the baseline method. It can be observed from Table 3 that randomly deleting one word significantly reduce the accuracy and F1-measure of sentiment classification. The accuracy decreases by up to 10.61% on STS-Gold using NB classifier in the N-grams model, and the F1-measure decreases by up to

TABLE 2. Gain/Loss in accuracy and Average F1-measure for removing URLs relative to not removing URLs Method using four classifiers for binary and 3-way sentiment classification on all datasets.

Gain/Loss (%)	Features Model	Classifier	SE-Twitter		SS-Twitter		STS-Gold		SemEval2014		STS-Test	
			binary	3-way	binary	3-way	binary	3-way	binary	3-way	binary	3-way
Accuracy	Prior polarity	LR	0	0	0	0	0	-	0	0	0	0
		NB	0	0.11	0	0	0	-	-0.13	0.22	0	0.2
		SVM	0	0	0	0	0	-	0.19	0.11	-0.28	-0.2
		RF	-0.50	0	-0.50	0.14	0.15	-	-0.38	0.11	0	0
	N-grams	LR	0	0	0	0	0	-	0	0	0	0
		NB	0	0	0	0	0	-	0	0	0	0
		SVM	0	0	0	0	0	-	0	0	0	0
		RF	0	0	0	0	0	-	0	0	0	0
F1-measure	Prior polarity	LR	0	0.11	0	-0.10	0	-	0	0	0	-0.18
		NB	0	0	0	0	0	-	-0.12	0.37	0	0.19
		SVM	0	0	0	0	0	-	0.19	0	-0.30	-0.19
		RF	-0.49	0	-0.50	0.19	0.16	-	-0.41	0.10	0	0.15
	N-grams	LR	0	0	0	0	0	-	0	0	0	0
		NB	0	0	0	0	0	-	0	0	0	0
		SVM	0	0	0	0	0	-	0	0	0	0
		RF	0	0	0	-0.1	0	-	0	0	0	0

TABLE 3. Gain/Loss in accuracy and Average F1- measure for removing URLs relative to the removal URLs replaced by removing one word Method using four classifiers for binary and 3-way sentiment classification on all datasets.

Gain/Loss (%)	Features Model	Classifier	SE-Twitter		SS-Twitter		STS-Gold		SemEval2014		STS-Test	
			binary	3-way	binary	3-way	binary	3-way	binary	3-way	binary	3-way
Accuracy	Prior polarity	LR	2.43	0	1.62	0	0.94	-	0	0	3.62	0
		NB	2.77	5.46	1.22	-1.26	-2.83	-	0	8.55	-0.28	1.81
		SVM	3.44	0.42	1.22	1.26	1.14	-	0.77	0.80	4.18	2.03
		RF	1.43	0	2.93	0.90	1.33	-	-0.26	-1.13	3.35	3.82
	N-grams	LR	2.86	0	2.36	0	0.30	-	0	0	3.06	0
		NB	2.44	0.83	1.84	6.03	10.61	-	1.98	1.09	1.11	0.40
		SVM	2.44	0.96	2.10	1.61	1.63	-	1.53	1.31	6.42	2.61
		RF	1.68	0.14	2.88	0.88	0.54	-	0.89	1.60	6.12	1.00
F1-measure	Prior polarity	LR	2.76	2.21	2.11	1.54	1.49	-	1.06	0.55	3.51	1.57
		NB	3.35	7.26	1.26	-3.24	-2.68	-	0	11.56	-1.02	2.23
		SVM	3.84	1.39	1.57	1.57	1.65	-	1.30	1.30	4.30	2.00
		RF	1.54	0.58	3.26	1.18	1.67	-	0	-1.23	3.27	3.93
	N-grams	LR	2.84	1.27	3.59	1.07	0.83	-	1.49	0	2.85	3.04
		NB	2.81	1.19	2.43	5.41	3.93	-	2.02	0	1.06	3.36
		SVM	2.49	1.86	4.07	1.34	2.13	-	2.20	0	6.36	7.52
		RF	1.68	0.82	5.30	1.16	0.68	-	2.01	0	6.13	1.05

TABLE 4. Gain/Loss in accuracy and Average F1- measure for removing stop words relative to not removing stop words Method using four classifiers for binary and 3-way sentiment classification on all datasets.

Gain/Loss (%)	Features Model	Classifier	SE-Twitter		SS-Twitter		STS-Gold		SemEval2014		STS-Test	
			binary	3-way	binary	3-way	binary	3-way	binary	3-way	binary	3-way
Accuracy	Prior polarity	LR	-0.42	0	-0.13	0	0.15	-	0	0	-1.11	0
		NB	1.43	-0.22	0	0.12	1.68	-	-0.77	-0.11	-0.56	0.20
		SVM	0	0	0	-0.31	0.40	-	0	-0.15	-0.27	0.60
		RF	-0.76	0	0.66	0.62	0.69	-	-0.32	-0.47	-0.84	0.20
	N-grams	LR	0	0	0	0	0	-	0	0	0	0
		NB	0	0	0	0	0	-	0	0	0	0
		SVM	0	0	0	0	0	-	0	0	0	0
		RF	0	0	0.35	0	-0.15	-	0	-0.15	0	-0.40
F1-measure	Prior polarity	LR	-0.41	0	-0.15	0	0.22	-	-0.17	0	-1.21	0
		NB	1.73	0	0	0.14	1.80	-	-0.69	-0.23	-0.76	0.24
		SVM	0.22	0	0	-0.24	0.30	-	-0.57	-0.22	-0.39	0.64
		RF	-0.50	-0.30	0.69	0.65	0.67	-	-0.43	-0.59	-0.86	0.25
	N-grams	LR	0	0	0	0	0	-	0	0	0	0
		NB	0	0	0	0	0	-	0	0	0	0
		SVM	0	0	0	0	0	-	0	0	0	0
		RF	0	0	0.36	0	-0.20	-	0	-0.19	0	-0.51

11.56% on the SemEval2014 using NB in the Prior polarity model. The accuracy and F1-measure of all classifiers reduce in the N-grams model. The performance of NB and RF fluctuate from −3.24% to 11.56%. From Table 2 and Table 3, it can be seen that the removal of URLs is an effective pre-processing method. The removal of URLs can reduce the vocabulary size effectively, whereas almost no impact on the performance of two types of classification tasks.

Table 4 is the result of a comparison of the performance of the baseline and the method that was applied other five pre-processing except for removal of stop words on all datasets. Table 4 shows that there is no effect of removing stop words on the performance of all classifiers in the N-grams model, except for RF. In the Prior polarity model, removing stop words causes fluctuation of the performance of all classifiers on the different datasets. On the STS-Test dataset, the perfor-

TABLE 5. Gain/Loss in accuracy and Average F1- measure for removing stop words relative to removing stop words replaced by removing one word Method using four classifiers for binary and 3-way sentiment classification on all datasets.

Gain/Loss (%)	Features Model	Classifier	SE-Twitter		SS-Twitter		STS-Gold		SemEval2014		STS-Test	
			binary	3-way	binary	3-way	binary	3-way	binary	3-way	binary	3-way
Accuracy	Prior polarity	LR	0	0	1.14	0	1.29	-	0	0	3.62	0
		NB	2.95	5.75	2.79	0.31	1.93	-	-0.77	-0.11	0.21	0.40
		SVM	0.92	0.90	0.83	1.12	1.43	-	0	-0.15	3.33	4.22
		RF	0.59	0	2.05	1.83	1.88	-	-0.19	-0.95	2.52	2.81
	N-grams	LR	1.85	0	1.66	0	0.89	-	0	0	0.83	0
		NB	1.68	0.44	2.01	0.21	18.04	-	0	0	3.65	1.00
		SVM	2.19	0.49	1.84	0.83	2.13	-	0	0	1.39	3.41
		RF	1.34	0.42	2.23	0.50	0.40	-	0	-0.15	2.77	1.00
F1-measure	Prior polarity	LR	0	2.15	1.65	1.70	1.77	-	-0.17	0	3.48	3.17
		NB	3.56	0.27	4.56	0.57	1.59	-	-0.69	-0.23	0.74	0.48
		SVM	1.11	1.85	1.14	1.74	1.95	-	-0.57	-0.22	3.40	4.40
		RF	0.79	0.58	2.38	1.85	2.14	-	-0.23	-0.74	2.48	2.86
	N-grams	LR	1.75	1.88	2.34	0.26	1.78	-	0	0	0.70	2.77
		NB	1.80	0.62	2.49	0.40	12.25	-	0	0	3.47	4.51
		SVM	2.22	1.36	2.67	0.55	3.34	-	0	0	1.30	4.15
		RF	1.38	1.34	2.93	0.86	0.87	-	0	-0.19	2.77	0.83

mance of all classifiers drop after removing stop words in the binary classification task, and improve after removing stop words in the 3-way classification task. In the Prior polarity model, the accuracy and average F1-measure of SVM fluctuates from −0.57% to 0.64%, and RF fluctuates from −0.86% to 0.69%, and NB fluctuates from −0.77% to 1.8% . The performance of NB classifier is more sensitive. In the Prior polarity model, the performance fluctuation of classifiers is due to the fact that the deleted stop words have a sentiment polarity value in the sentiment dictionary.

Table 5 is the result of a comparison of the performance of baseline and the removal stop words replaced by random deletion of one word in the baseline method on all datasets. Table 5 indicates that random deleting one word reduce the accuracy and F1-measure of all classifiers on all datasets except for SemEval2014 dataset. On the SemEval2014 dataset, there is improvement in the performance of sentiment classification for the Prior polarity model, but there is little change in the N-grams model. From Table 4 and Table 5, it can be observed that removing stop words is an effective pre-processing method while using the N-gram model. The removal of stop words can reduce the vocabulary size in the N-gram model effectively. In the Prior polarity model, the length of the feature vector is not affected by the word space. Deleting stop words does not reduce the feature vector space.

Table 6 is the result of a comparison of the performance of the baseline and the method that was applied other five pre-processing method except for removal of numbers on all datasets. From Table 6, it can be observed that removing numbers does not affect the performance of all classifiers on all datasets in the Prior polarity model, except RF. In the Prior polarity model, the performance of RF fluctuates from −0.18% to 0.3%, the fluctuation range is limited to 0.3%. One factor of the fluctuating of RF performance is the initial random value of the classifiers. The performance of SVM obtains a small improvement after removing numbers on all datasets in the N-grams model. The accuracy and F1-measure of LR classifier change do not exceed 0.5% in the N-grams model. In the N-grams model, the performance of NB and RF are fluctuation in two types of classification task. The maximum increase of accuracy and F1-measure of NB is 0.25% and 0.28% respectively on the SE-Twitter dataset, and the maximum reduction of accuracy and F1-measure of NB is −2.23% and −2.95% respectively on the STS-Gold dataset in the binary classification task. The maximum reduction of accuracy and F1-measure of RF is −0.83% on the STS-Test dataset and −1.01% on the STS-Gold dataset respectively in the binary classification task, and the maximum increase of accuracy and F1-measure of RF is 0.17% and 0.19% respectively on the SS-Twitter in the 3-way classification task.

Table 7 is the result of a comparison of the performance of the baseline and the removal number replaced by deletion one word randomly in the baseline C-Method on all datasets. Table 7 indicates that random deleting one word reduces the accuracy and F1-measure of sentiment classification. From Table 6 and Table 7, it can be observed that removing numbers almost does not impact on the performance of two types of classification tasks. Removing number is an effective pre-processing method.

Table 8 is the result of a comparison of the performance of the baseline and the method applied other five pre-processing except for reverting words that contain repeated letter on all datasets. It can be observed from table 8 that the performance

TABLE 6. Gain/Loss in accuracy and Average F1- measure for removing numbers relative to not removing numbers Method using four classifiers for binary and 3-way sentiment classification on all datasets.

Gain/Loss (%)	Features Model	Classifier	SE-Twitter		SS-Twitter		STS-Gold		SemEval2014		STS-Test	
			binary	3-way	binary	3-way	binary	3-way	binary	3-way	binary	3-way
Accuracy	Prior polarity	LR	0	0	0	0	0	-	0	0	0	0
		NB	0	0	0	0	0	-	0	0	0	0
		SVM	0	0	0	0	0	-	0	0	0	0
		RF	0	0	0	0	0.30	-	-0.13	-0.18	0	0
	N-grams	LR	0.34	0	-0.18	0	-0.30	-	0	0	-0.49	0
		NB	0.25	0	0	0	-2.23	-	0.13	0.11	-0.29	0.20
		SVM	0.34	0	0.31	0.33	0.40	-	0.26	0	0.28	0.80
		RF	0	-0.16	-0.44	0.17	-0.74	-	0.13	0	-0.83	0
F1-measure	Prior polarity	LR	0	0	0	0	0	-	0	0	0	0
		NB	0	0	0	0	0	-	0	0	0	0
		SVM	0	0	0	0	0	-	0	0	0	0
		RF	0	-0.16	0	0	0.20	-	-0.20	-0.18	0	0
	N-grams	LR	0.38	0	-0.15	-0.11	-0.34	-	0	0.17	-0.48	0.13
		NB	0.28	0	0	0	-2.95	-	0.13	0.22	-0.51	0.15
		SVM	0.36	0	0.33	0	0.50	-	0.23	0	0.28	0.69
		RF	0	-0.42	-1.0	0.19	-1.01	-	0.18	0	-0.81	-0.18

TABLE 7. Gain/Loss in accuracy and Average F1- measure for removing numbers relative to removing numbers replaced by removing one word Method using four classifiers for binary and 3-way sentiment classification on all datasets.

Gain/Loss (%)	Features Model	Classifier	SE-Twitter		SS-Twitter		STS-Gold		SemEval2014		STS-Test	
			binary	3-way	binary	3-way	binary	3-way	binary	3-way	binary	3-way
Accuracy	Prior polarity	LR	1.51	0	1.49	0	1.04	-	0	0	6.40	0
		NB	4.22	-4.91	6.78	0.52	-2.47	-	0.89	0	1.67	1.01
		SVM	2.09	0.85	1.40	1.26	1.14	-	1.91	0.44	1.37	2.80
		RF	0.76	0.23	2.14	0.83	1.92	-	0.26	-0.15	2.79	4.61
	N-grams	LR	2.76	0	1.92	0	0.59	-	0	0	2.50	3.01
		NB	2.68	0.54	1.53	0.62	-2.13	-	2.74	1.27	0	0
		SVM	2.85	0.53	2.14	1.52	1.93	-	1.21	0.98	3.34	-0.39
		RF	2.18	0.47	1.97	1.45	0.55	-	0.38	1.38	3.61	4.62
F1-measure	Prior polarity	LR	1.59	3.11	1.92	1.66	1.66	-	1.22	1.11	6.33	3.00
		NB	5.36	-3.29	6.54	0.68	-2.61	-	1.37	-0.89	2.11	1.18
		SVM	2.59	1.97	1.78	1.76	1.93	-	4.20	0.63	1.47	3.08
		RF	0.82	0.68	2.23	1.34	1.99	-	0.72	-0.51	2.83	4.53
	N-grams	LR	2.73	2.42	1.81	1.04	1.26	-	0.74	1.83	2.46	1.56
		NB	3.01	0.77	1.82	1.11	-3.10	-	2.83	2.05	0.20	-1.90
		SVM	2.79	1.53	3.85	1.41	2.70	-	1.64	1.26	3.39	8.48
		RF	2.22	1.11	3.40	1.81	1.02	-	1.15	1.75	3.77	2.26

of all classifiers does not change except for RF in the 3-way classification task, and the NB is more sensitive to reverting repeated letters than other classifiers in the Prior polarity model. In the N-grams model, after reverting repeated letters, the performance of LR and RF obtain increase on all datasets in the binary classification task, and the performance of NB drops on the STS-Gold and SemEval2014 datasets in the binary classification task. In the binary classification task, the performance of SVM is fluctuation, after reverting repeated letters. The maximum increase of the accuracy and F1-measure of SVM is 0.79% and 1.4% respectively on the SS-Twitter dataset in the Prior polarity model, and the maximum reduction of the accuracy and F1-measure of SVM is −0.76% and −0.77% respectively on the SE-Twitter dataset in the N-grams model.

Table 9 is the result of a comparison of the performance of the baseline and the method applied other five pre-processing except for expanding acronym on all datasets.

TABLE 8. Gain/Loss in accuracy and Average F1- measure for reverting repetition relative to not reverting repetition Method using four classifiers for binary and 3-way sentiment classification on all datasets.

Gain/Loss (%)	Features Model	Classifier	SE-Twitter		SS-Twitter		STS-Gold		SemEval2014		STS-Test	
			binary	3-way	binary	3-way	binary	3-way	binary	3-way	binary	3-way
Accuracy	Prior polarity	LR	-1.01	0	0	0	0.20	-	0	0	0.56	0
		NB	-12.11	0	3.41	0	7.57	-	1.85	0	0	0
		SVM	-0.76	0	-0.35	0	0.25	-	0.19	0	-0.56	0
		RF	-1.25	-0.24	0.26	-0.24	0.35	-	-0.13	-0.11	-0.28	0.20
	N-grams	LR	0.25	0	0.17	0	0	-	0	0	0.28	0
		NB	0	0	0	0	-0.25	-	-0.13	0	0	0
		SVM	-0.33	0	0.79	0	0	-	0	0	-0.28	0
		RF	0.50	0	1.22	0	0.10	-	0	0	0.83	-0.20
F1-measure	Prior polarity	LR	-1.10	0	-0.10	0	0.27	-	-0.19	0	0.54	0
		NB	-12.77	0	3.25	0	9.32	-	12.52	0	0	0
		SVM	-0.77	0	-0.35	0	0.33	-	0	0	-0.59	0
		RF	-1.27	-0.14	0.25	-0.28	0.29	-	-0.25	0	-0.23	0.17
	N-grams	LR	0.23	0	0.17	0	0	-	0.13	0	0.26	0
		NB	0	0	0	0	-0.19	-	-0.13	0	0	0
		SVM	-0.41	0	1.40	0	0	-	0	0	-0.28	0
		RF	0.48	0	1.77	-0.10	0.24	-	0	0	0.82	-0.25

TABLE 9. Gain/Loss in accuracy and Average F1- measure for expanding acronym relative to not expanding acronym Method using four classifiers for binary and 3-way sentiment classification on all datasets.

Gain/Loss (%)	Features Model	Classifier	SE-Twitter		SS-Twitter		STS-Gold		SemEval2014		STS-Test	
			binary	3-way	binary	3-way	binary	3-way	binary	3-way	binary	3-way
Accuracy	Prior polarity	LR	0	0	0.79	0	1.73	-	0	0	1.95	0
		NB	-0.34	0	0.92	0	-8.06	-	0.19	0	0.28	0
		SVM	0.76	0	0.17	0	1.78	-	1.53	0	2.51	0
		RF	-1.34	-0.13	2.36	0	0.99	-	0	-0.15	-0.57	-0.20
	N-grams	LR	1.35	0	2.27	0	0.89	-	0	0	1.11	0
		NB	1.09	0	0.44	0	6.85	-	-1.02	0	-1.40	0
		SVM	0.76	0	1.53	0	1.78	-	0.70	0	1.67	0
		RF	0.59	0	1.79	-0.12	0.89	-	-0.19	-0.15	0	-0.20
F1-measure	Prior polarity	LR	0	0	1.02	0	2.44	-	1.01	0	2.06	0
		NB	-0.42	0	0.66	0	-7.61	-	-0.13	0	0.63	0
		SVM	0.78	0	0.19	0	2.56	-	3.16	0	2.88	0
		RF	-1.40	-0.23	2.82	0	1.28	-	0.48	-0.17	-0.21	-0.18
	N-grams	LR	1.32	0	2.86	0	2.03	-	0.14	0	1.13	0
		NB	1.33	0	0.38	0	6.08	-	-1.03	0	-1.45	0
		SVM	0.88	0	3.54	0	2.99	-	0.63	0	1.72	0
		RF	0.59	-0.16	3.94	-0.14	1.91	-	-0.17	-0.19	0	-0.26

Table 9 shows that the performance of all classifiers do not change except for RF in the 3-way classification task and the performance of LR and SVM is improvement in the two feature models after expanding acronyms in two types of task. The accuracy and F1-measure of NB reach the maximum of 6.85% and 6.08% improvement in the N-grams model and drop 8.06% and 7.61% in the prior polarity model on the STS-Gold dataset, respectively. In the binary classification task, expanding acronyms promotes the accuracy and F1-measure of RF on the SE-Twitter, SS-Twitter and STS-Gold datasets but decreases these values by 0.19% and 0.17% on the SemEval2014 dataset in the N-grams model, and decreases by 1.34% and 1.4% on the SE-Twitter, and by 0.57% and 0.21% on STS-Test datasets in the Prior polarity model, respectively.

Table 10 is a comparison of the performance of the baseline and the method that was applied other five pre-processing method except for replacement of negative mentions on all

TABLE 10. Gain/Loss in accuracy and Average F1- measure for replacing negation relative to not replacing negation Method using four classifiers for binary and 3-way sentiment classification on all datasets.

Gain/Loss (%)	Features Model	Classifier	SE-Twitter		SS-Twitter		STS-Gold		SemEval2014		STS-Test	
			binary	3-way	binary	3-way	binary	3-way	binary	3-way	binary	3-way
Accuracy	Prior polarity	LR	-1.44	0	0.79	0	2.08	-	0	0	1.69	0
		NB	-0.17	2.25	1.79	0.14	-0.94	-	1.72	0.40	0.28	1.81
		SVM	0.58	0.1	3.58	0.62	2.13	-	2.10	0.69	0.83	0.61
		RF	-2.02	-0.65	2.36	1.26	2.87	-	2.55	0	0	0.40
	N-grams	LR	1.35	0	0.44	0	0.45	-	0	0	2.23	0
		NB	2.28	0	0.44	0.52	-3.12	-	0.89	0.95	0.83	1.01
		SVM	1.52	0.10	1.84	1.80	2.92	-	8.23	1.49	0	5.23
		RF	1.43	0.19	1.14	2.45	1.73	-	2.23	1.64	1.96	1.81
F1-measure	Prior polarity	LR	-1.63	2.21	0.50	0.50	2.62	-	4.11	2.13	1.73	0.89
		NB	-0.23	7.26	1.84	0.56	-0.76	-	3.52	1.25	0	2.36
		SVM	0.71	1.39	4.77	1.80	2.38	-	4.80	2.02	1.16	0.96
		RF	-2.12	0.58	2.64	2.22	2.78	-	4.30	0.58	0	0.72
	N-grams	LR	1.57	0.30	0	5.01	1.09	-	0.11	2.29	2.07	6.67
		NB	2.59	0	0.48	0.67	-4.41	-	0.91	0.90	1.31	1.34
		SVM	1.68	0.56	1.13	5.45	3.39	-	1.26	7.78	0	6.23
		RF	1.63	0	0.61	6.69	1.99	-	10.21	7.23	1.87	2.35

datasets. Table 10 shows that there is a significant increase in the accuracy and F1-measure of all classifiers after replacing negation in the N-grams model on all five datasets in two types of classification task, except for NB on the STS-Gold dataset. After replacing negation, the maximum improvement of accuracy is 8.23% using the SVM classifier and the improvement of the F1-measure is 10.21% using the RF classifier on the SemEval2014 dataset. In the Prior polarity model, the performance of all classifiers improves after replacing negation on the SS-Twitter, SemEval2014 and STS-Test datasets, and the performance of NB drop on the STS-Gold dataset. However, the performance of LR, NB and RF drops on the SE-Twitter dataset in the Prior polarity model in the binary task. The performance of SVM increases on all datasets in two types of classification task.

V. DISCUSSION

The experimental results show that removing URLs barely affects the performance of classifiers in the two feature models on all datasets. This indicates that URLs do not contain useful information for sentiment classification. Table 4 shows that there is few effects on the performance of classifiers in the N-grams model before and after removing stop words. One of the reasons might be that stop words appear in tweets frequently. In the Prior polarity model, removing stop words leads to the fluctuation of classifier performance because a stop words contains different sentiment polarity. The results in Table 4 suggest that it is necessary to remove stop words for sentiment classification. It can be observed from Table 6 that removing numbers has no effect on the accuracy of sentiment classification in the Prior polarity model because the numbers are neutral. In the N-grams model, the removal of numbers causes fluctuation of classifier performance, except for SVM. The performance of SVM improves on all datasets after removing numbers. Therefore, removing numbers is useful to improve the performance of sentiment classification using SVM. The effect of removing repeated letters on the performance of classifiers is different on each dataset, which suggests that removing repeated letters influences the polarity and semantic features of words in tweets. Expanding acronyms improves the performance of classifiers on most datasets. Expanding acronyms to their original words is a more formal expression than an acronym. The results in Table 10 show that the performance of classifiers increases after replacing negation on all datasets in most cases because negation contains important sentiment polarity features. In the N-grams model, removing URLs and removing stop words reduce the vocabulary size, and there is no change in the performance of all classifiers, but removing numbers affects the performance. In the Prior polarity model, removing numbers does not affect the performance of classifiers. The performance of SVM increases after replacing negation and expanding acronym, therefore, replacing negation and expanding acronym are effective pre-processing method while using SVM classifier. Reverting words that contain repeated letter causes performance to fluctuate, therefore, the pre-processing method is not recommended. From Table 3, Table 5, and Table 7, it can be observed that the random deletion of words causes a significant decline in the performance of the classification because the randomly deleted word may be a missing key word, causing decision polarity or damage to the sentence semantic relationship.

The experimental results show that the same pre-processing method affects the performance of sentiment

classifiers similarly, whereas the NB and RF classifiers are more sensitive than LR and SVM classifiers. One factor that may affect the results of sentiment classification is the choice of the sentiment classifier and the features used for classifier training.

In the future, we will continue our evaluation using different stoplists and acronym dictionaries and will investigate the reasons for the fluctuation of sentiment classification performance using different classifiers on various datasets.

VI. CONCLUSION

This paper studies that six different pre-processing methods affect sentiment polarity classification in the Twitter. We conduct a series of experiments using four classifiers to verify the effectiveness of several pre-processing methods on five Twitter datasets. Experimental results indicate that the removal of URLs, the removal of stop words and the removal of numbers minimally affect the performance of classifiers; furthermore, replacing negation and expanding acronyms can improve the classification accuracy. Therefore, removing stop words, numbers, and URLs is appropriate to reduce noise but does not affect performance. Replacing negation is effective for sentiment alalysis. We select appropriate pre-processing methods and feature models for different classifiers for the Twitter sentiment classification task.

REFERENCES

[1] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!" in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, 2011, pp. 538–541.

[2] D. Terrana, A. Augello, and G. Pilato, "Automatic unsupervised polarity detection on a Twitter data stream," in *Proc. IEEE Int. Conf. Semantic Comput.*, Newport Beach, CA, USA, Sep. 2014, pp. 128–134.

[3] H. Saif, Y. He, M. Fernandez, and H. Alani, "Semantic patterns for sentiment analysis of Twitter," in *Proc. 13th Int. Semantic Web Conf.*, Apr. 2014, pp. 324–340.

[4] H. Saif, M. Fernandez, Y. He, and H. Alani, "On stopwords, filtering and data sparsity for sentiment analysis of Twitter," in *Proc. 9th Lang. Resour. Eval. Conf. (LREC)*, Reykjavik, Iceland, 2014, pp. 80–81.

[5] H. Saif, Y. He, and H. Alani, "Alleviating data sparsity for Twitter sentiment analysis," in *Proc. CEUR Workshop*, Sep. 2012, pp. 2–9.

[6] H. G. Yoon, H. Kim, C. O. Kim, and M. Song, "Opinion polarity detection in Twitter data combining shrinkage regression and topic modeling," *J. Informetrics*, vol. 10, no. 2, pp. 634–644, 2016.

[7] F. H. Khan, U. Qamar, and S. Bashir, "SentiMI: Introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection," *Appl. Soft Comput.*, vol. 39, pp. 140–153, Apr. 2016.

[8] A. Agarwal, B. Xie, and I. Vovsha, "Sentiment analysis of Twitter data," in *Proc. Workshop Lang. Social Media, Assoc. Comput. Linguistics*, 2011, pp. 30–38.

[9] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia Comput. Sci.*, vol. 17, pp. 26–32, Sep. 2014.

[10] Y. Bao, C. Quan, L. Wang, and F. Ren, "The role of pre-processing in Twitter sentiment analysis," in *Proc. 10th Int. Conf. (ICIC)*, Taiyuan, China, pp. 615–624, Apr. 2014.

[11] Z. Jianqiang, "Pre-processing boosting Twitter sentiment analysis?" in *Proc. IEEE Int. Conf. Smart City/SocialCom/SustainCom (SmartCity)*, Sep. 2015, pp. 748–753.

[12] C. J. V. Rijsbergen, "Information retrieval," in *Butterworth-Heinemann*, 2nd ed. Newton, MA, USA, 1979.

[13] C. Fox, "Information retrieval data structures and algorithms," in *Lexical Analysis and Stoplists*, Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1992, pp. 102–130.

[14] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," Stanford, Stanford, CA, USA, Project Rep. CS224N, 2009.

[15] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proc. LREC*, vol. 10. 2010, pp. 1320–1326.

[16] X. Hu, J. Tang, H. Gao, and H. Liu, "Unsupervised sentiment analysis with emotional signals," in *Proc. 22nd World Wide Web Conf.*, 2013, pp. 607–618.

[17] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social Web," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 1, pp. 163–173, 2012.

[18] G. Paltoglou and M. Thelwall, "Twitter, myspace, digg unsupervised sentiment analysis in social media," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 4, pp. 1–19, 2012.

[19] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proc. 7th Conf. Int. Lang. Resour. Eval.*, Valletta, Malta, 2010, pp. 2200–2204.

[20] A. Bakliwal, P. Arora, S. Madhappan, N. Kapre, M. Singh, and V. Varma, "Mining sentiments from tweets," in *Proc. 3rd Workshop Comput. Approaches Subjectivity Sentiment Anal., Assoc. Comput. Linguistics*, Jeju, South Korea, 2012, pp. 11–18.

[21] H. Saif, M. Fern, and Y. He, "Evaluation datasets for Twitter sentiment analysis: A survey and a new dataset, the STS-gold ," in *Proc. 1st ESSEM Workshop*. Turin, Italy, 2013, pp. 21–26.

[22] S. Narr, M. Hulfenhaus, and S. Albayrak, "Language-independent Twitter sentiment analysis," Knowledge Discovery and Machine Learning (KDML), LWA, 2012, pp. 12–14.

ZHAO JIANQIANG is currently pursuing the Ph.D. degree with Xi'an Jiaotong University, Xi'an, China. His research interests include social network and web mining.

GUI XIAOLIN received the B.Sc. degree in computer from the Xi'an Jiaotong University of China (XJTU) and the M.Sc. and Ph.D. degrees in computer science from XJTU in 1993 and 2001, respectively. Since 1988, he has been with XJTU, as an Active Researcher in network computing, network security, and wireless networks, where he is currently a Professor and a Director of the Key Laboratory of Computer Network. His recent research covers secure computation of open network systems, including grid, P2P, and cloud, dynamic trust management theory, and development on community network.