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

Naveen Ravipati
Department of Computer Engineering
San Jose State University
San Jose, CA, United States
naveen.ravipati@sjsu.edu

***Abstract*— Sentiment analysis is the process of determining the public opinion based on the data. Here, sentiment analysis is carried out on twitter dataset to classify the data as neutral or negative. The first step of machine learning algorithm is pre-processing of the data. Then, the pre-processed data will be used to extract the features. The effect of each text pre-processing step on the performance of the binary classification is verified using two models and four classifiers. There is a significant improvement in the performance when the pre-processing methods such as expanding the contractions and reverting the repeated letters are employed but barely changes when removing URLs, removing numbers or stop words are used. Random Forest and Naive Bayes are more responsive than Logistic Regression and support vector machine classifiers when the pre-processing methods are applied.**

Keywords—Sentiment Analysis, Accuracy, F1-Score, Comparison.

I. INTRODUCTION

Twitter is an American online news and social networking service on which users post and interact with messages known as "tweets". Sentiment analysis is the process of detecting the opinionated content from the text data and it can determine text polarity. Sentimental analysis can be carried out on these tweets to obtain the public opinion. This information can be used by political and economic organizations to enhance their reputation.

Tweets are usually composed of noisy, improper and poorly structured sentences. Before the feature extraction, a set of pre-processing steps (e.g., removing stop words, removing URLs, expanding the contractions) are done to remove the noise from the data. Pre-processing is already done in a lot of machine learning approaches [1]-[2][3][4]. However, only a few studies are done on the effect of pre-processing methods on the performance of the twitter sentiment analysis classification. This paper evaluated the impact of various pre-processing methods on twitter sentiment classification. Two feature models and four machine learning classifiers are used to identify the tweet sentiment polarity.

II. RELATED WORK AND BACKGROUND

In the previous papers [1]-[2][3][4][5][6][7][8] pre-processing is done on tweets to reduce the noise. The assumption is that pre-processing reduces noise in the data, and it should improve the performance of the classification task. Haddi et al. [9] explored the role of pre-processing in movie reviews sentiment analysis and the results show that the accuracy of sentiment classification can be improved significantly by using appropriate features. Saif et al. [4] projected the effect of stop words removal on classification of tweets and assessed the impact of removing stop words by monitoring level of data sparsity, size of the classifier feature space and the performance of the classification. Saif et al. [5] found that pre-processing led to a significant reduction and the dictionary size was reduced to 62%. However, the performance of the classification task was not discussed. Bao et al. [10] explored the effect of pre-processing on the twitter sentiment classification. The results on the Stanford Twitter Dataset show that the performance increases when the URL features reservation, negative transformation and normalizing repeated letters but decreases when stemming and lemmatization are employed. Zhao [11] has done similar analysis and his results show that accuracy increases after expanding acronym and replacing negative mentions but hardly changes when removal of URL, removal of numbers and stop words are applied.

From the literature review, there is no in-depth analysis of impact of pre-processing on Twitter sentiment classification. To cover this gap, this article focuses on evaluating the effects of text pre-processing on Twitter sentiment classification using two feature models and four machine learning classifiers.

III. PRE-PROCESSING ANALYSIS SETUP

To assess the effect of various pre-planning method, five pre-processing methods are applied to sentiment classification using four different classifiers.

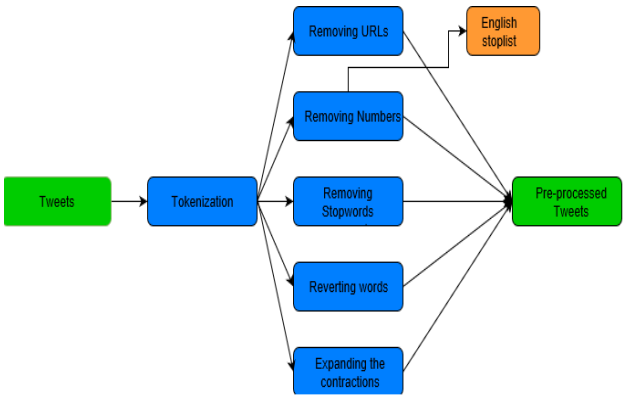

Fig.1. Twitter text pre-processing process

A. *Pre-Processing*

Pre-processing methods used in this paper are illustrated in the Fig.1 and are as follows:

- Expanding the contractions. Tweets usually consists of contractions. Contractions usually play a significant role in the sentiment analysis. Here the process of expanding the contractions is transforming won't, can't, n't, 've, 'm into willnot, cannot, not, have and am.
- Removing URLs in the tweets. In most of the cases, URLs don't contain information regarding the sentiment of the tweet. Hence, this removal will improve the quality of the content in the tweet.
- Reverting words to their original form. Tweets usually consists of words that are an extension of the words that are in the English dictionary. These words should be transformed to their original form. Reverting back to their original form can reduce the feature size. For example, "coooool" is replaced by "coool".
- Removing stop words. Stop words usually refer to the most commonly used words in English. These words don't ay information to the sentimental analysis. These words can be excluded from the tweets. One way of the stop words removal is the usage of pre-compiled lists.
- Removal of numbers. To refine the tweets, numbers are usually removed.

B. *Feature Models*

- Word n-grams Feature Model

Word n-grams features are the simplest feature for Twitters sentiment analysis. Word n-grams are the state of the art for the sentiment analysis. In this paper, Word unigram is the feature model.

- Tf-idf Feature Model

Tf-idf stands for term frequency-inverse frequency domain, is a numerical statistic that is intended to reflect how important a word is to document in a collection or corpus. Tf-idf is a feature model that is used in the sentiment analysis.

C. *Twitter Sentiment Classifiers*

To assess the effect of pre-processing on sentiment classification, 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*

A binary classification of classifying the data into negative and neutral is performed on the tweets data. The classification is performed by using SVM, NB, LR and RF classifiers. The data is pre-processed using five pre-processing methods and the features are extracted using the two models and the final data is passed onto the classifiers to perform the classification.

The performance of the sentiment classification is evaluated using accuracy and F1- measure. The gain or loss of accuracy and F1-measure are calculated based on the impact of removal of certain step in the text pre-processing.

E. *Dataset*

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 2102 negative and 1101 neutrals tweets. Although the Stanford test set is relatively small, it has been widely used in the literature [5], [20] for different evaluation tasks.

IV. Experimental Results

In this section, the baseline results are compared to the results obtained after applying several pre-processing methods. The baseline consists of the results in which all the pre-processing methods are applied. The effect of the pre-processing method is calculated by observing the increase or decrease in the performance of accuracy or F1-measure. The accuracy improvement is calculated as follows:

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

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

$$F1_{improvement} = \text{Average } F1_{baseline} - \text{Average } F1_{compared}$$

TABLE I. GAIN/LOSS IN ACCURACY AND AVERAGE F1-MEASURE FOR REMOVING URLS RELATIVE TO NOT REMOVING URLS METHOD USING FOUR CLASSIFIERS FOR BINARY SENTIMENT CLASSIFICATION

Evaluation Criteria	Classifier	Performance Improvement	
		N-grams	*Tf-idf*
Accuracy	LR	-1.3	-0.7
	NB	0.3	0
	SVM	-0.7	-1.2
	RF	0.6	-0.6
F1-measure	LR	0.2	-0.5
	NB	0	-0.3
	SVM	0	-0.7
	RF	-1	0

Table I 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 four pre-processing method except for removal of URLs. It can be observed from Table I that the performance of every classifier in the N-grams model and Tf-idf does not change after removing the URLs.

TABLE II. GAIN/LOSS IN ACCURACY AND AVERAGE F1-MEASURE FOR REMOVING STOP WORDS RELATIVE TO NOT REMOVING STOP WORDS METHOD USING FOUR CLASSIFIERS FOR BINARY SENTIMENT CLASSIFICATION

Evaluation Criteria	Classifier	Performance Improvement	
		N-grams	*Tf-idf*
Accuracy	LR	0.7	-0.3
	NB	0	-0.6
	SVM	-0.7	-0.2
	RF	-0.7	-0.3
F1-measure	LR	0.2	-0.1
	NB	0	-0.4
	SVM	0	0.4
	RF	-1	-0.2

Table II reports the effect of the removal of stop words on classification performance and is the result of a comparison of the performance of the baseline and the method that was applied other four pre-processing method except for removal of stop words. It can be observed from Table II that the performance of every classifier in the N-grams model and Tf-idf does not change after removing the stop words.

TABLE III. GAIN/LOSS IN ACCURACY AND AVERAGE F1-MEASURE FOR REMOVING NUMBERS RELATIVE TO NOT REMOVING NUMBERS METHOD USING FOUR CLASSIFIERS FOR BINARY SENTIMENT CLASSIFICATION

Evaluation Criteria	Classifier	Performance Improvement	
		N-grams	*Tf-idf*
Accuracy	LR	1	-0.3
	NB	-0.3	0.7
	SVM	-0.3	-0.3
	RF	0.3	-0.6
F1-measure	LR	0	0
	NB	0	0.3
	SVM	-0.3	0.3
	RF	-0.5	0.6

Table III reports the effect of the removal of numbers on classification performance and is the result of a comparison of the performance of the baseline and the method that was applied other four pre-processing method except for removal of numbers. It can be observed from Table III that the performance of every classifier in the N-grams model and Tf-idf does not change after removing the numbers.

TABLE IV. GAIN/LOSS IN ACCURACY AND AVERAGE F1-MEASURE FOR NOT REVERTING WORDS TO ORIGINAL FORM RELATIVE TO REVERTING WORDS METHOD USING FOUR CLASSIFIERS FOR BINARY SENTIMENT CLASSIFICATION

Evaluation Criteria	Classifier	Performance Improvement	
		N-grams	*Tf-idf*
Accuracy	LR	1.7	0
	NB	-0.3	0.4
	SVM	-2.3	-1.6
	RF	2.3	3
F1-measure	LR	0	0.5
	NB	-0.2	0.3
	SVM	0.8	1.1
	RF	0.2	-0.2

Table IV shows the effect of not reverting the words to their original form to reverting words. It can be seen that there is a peak change of 3% for RF classifier using Tf-idf model. This result indicates that reverting words has a significant impact on the performance of the twitter sentimental analysis task.

TABLE V. GAIN/LOSS IN ACCURACY AND AVERAGE F1-MEASURE FOR NOT EXPANDING THE CONTRACTIONS RELATIVE TO EXPANDING THE CONTRACTIONS METHOD USING FOUR CLASSIFIERS FOR BINARY SENTIMENT CLASSIFICATION

Evaluation Criteria	Classifier	Performance Improvement	
		N-grams	*Tf-idf*
Accuracy	LR	-0.6	-2
	NB	-0.3	0.4
	SVM	-1.7	-4.6
	RF	0	-2.6
F1-measure	LR	0.3	-0.7
	NB	-1.3	-0.1
	SVM	1.3	2.3
	RF	-0.2	-0.2

Table V shows the effect of not expanding the contractions to expanding the contractions. It can be seen that there is a peak change of -4.6% for SVM classifier using tf-idf model. This result indicates that reverting words has a significant impact on the performance of the twitter sentimental analysis task.

V. CONCLUSION

This paper evaluates the impact of five pre-processing methods on the performance of binary classification in the twitter sentimental analysis task. The effectiveness is verified using four different classifiers. Experimental results indicate that the removal of stop words, removal of URLs and the removal of numbers minimally affect the performance of classifiers but expanding the contractions and reverting the words to original form can improve the classification accuracy. Therefore, removing stop words, numbers, and

URLs is appropriate for noise removal but does not impact performance. Expanding the contractions and reverting the words to original form effects the performance of classification in the twitter sentimental analysis task.

REFERENCES

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

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

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

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

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

[6] H. G. Yoon, H. Kim, C. O. Kim, M. Song, "Opinion polarity detection in Twitter data combining shrinkage regression and topic modeling", *J. Informetrics*, vol. 10, pp. 634-644, 2016.

[7] F. H. Khan, U. Qamar, 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, I. Vovsha, "Sentiment analysis of Twitter data", *Proc. Workshop Lang. Social Media Assoc. Comput. Linguistics*, pp. 30-38, 2011.

[9] E. Haddi, X. Liu, 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, F. Ren, "The role of pre-processing in Twitter sentiment analysis", *Proc. 10th Int. Conf. (ICIC)*, pp. 615-624, Apr. 2014.

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

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

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

[14] A. Go, R. Bhayani, L. Huang, "Twitter sentiment classification using distant supervision", 2009.

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

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

[17] M. Thelwall, K. Buckley, 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, 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, F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining", *Proc. 7th Conf. Int. Lang. Resour. Eval.*, pp. 2200-2204, 2010.

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

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

[22] S. Narr, M. Hulfenhaus, S. Albayrak, "Language-independent Twitter sentiment analysis", pp. 12-14, 2012.