Title: Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis Authors: ZHAO JIANQIANG and GUI XIAOLIN

Feature models used

1. Ngrams
2. Prior polarity

SenScore(w) = PMI(w, pos) − PMI(w, neg)

a positive SenScore (w) suggests a stronger association of word w with positive sentiment and vice versa

PMI - pointwise mutual information

Sentiment Classifiers used

1. SVM
2. Naives Bayes
3. Logistic Regression
4. Random Forest

Method used :

1. six preprocessing steps are considered.

* Replacing negative mentions
* Removing URL
* Reverting words
* Removing numbers
* Removing stop words
* Expanding acronyms.

1. All done separately on a baseline method(C method) for both ngrams and PPS feature models.
2. Then change in performance is recorded over different classifiers.
3. 2 way( + - ) and 3 way (+,-,neutral) classification done over 5 datasets.

Results:

* NB and RF classifiers are more sensitive than LR and SVM classifiers.
* Removing stop words leads to the fluctuation of classifier performance on prior polarity models because the stop words contain different sentiment polarity. It has no effect on Ngrams models.
* removing repeated letters influences the polarity and semantic features of words in tweets,increasing performance.
* Removing numbers mostly has no effect except in SVM where there is slight improvement in performance.
* Reverting words with repeated letters causes fluctuation in performance.
* **Removing stop words, numbers, and URLs is appropriate to reduce noise but does not affect performance.**
* **Replacing negation and expanding acronyms can improve the classification accuracy.**