TWITTER SENTIMENT ANALYSIS USING CLASSIFICATION AND REGRESSION TREE (CART) AND RANDOM FOREST

LCR Paquit, VB Calag, CM Garillos, AR Mesa

Department of Math, Physics and Computer Science, College of Science and Mathematics,
University of the Philippines in Mindanao, Bago Oshiro, Mintal, Davao City

INTRODUCTION

The main objective of this study was to compare the performance and accuracy of sentiment analysis of tweets with distant supervision using Classification and Regression Tree (CART) and Random Forest to the results of sentiment analysis using Naïve Bayes, Maximum Entropy, and Support Vector Machine (SVM). The study implements pre-processing procedures on the training and testing data and apply CART and Random Forest on each subset of the tweets for classification. The classifiers were interpreted and evaluated under different training set sizes. The best performing classifiers were tested to classify a real Twitter dataset.

MATERIALS AND METHODS

A stratified sample of sizes 4,000, 8,000 and 12,000 tweets were extracted and used as training sets for this study. A test set of manually annotated tweets consisted of 177 negative and 182 positive tweets. A real Twitter dataset was also extracted using a Twitter API. The data was collected on the 31st of March 2016 using the keyword “Duterte”.

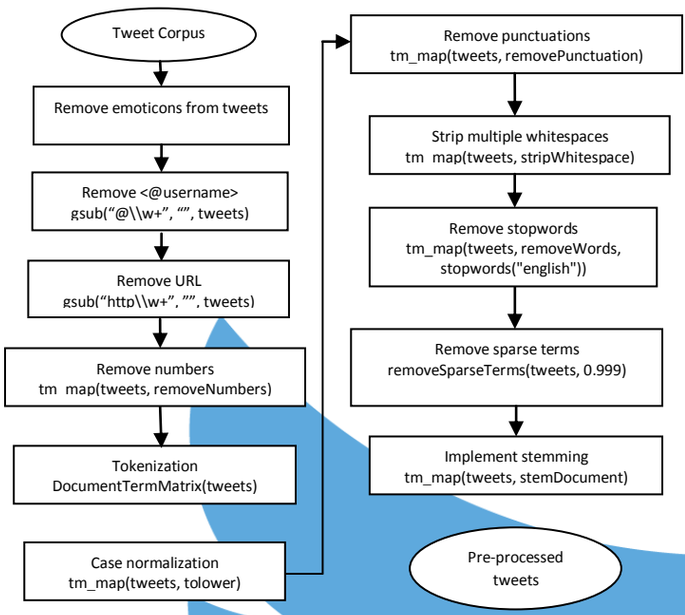

Fig. 1. Pre-processing tweets in R.

RESULTS AND DISCUSSION

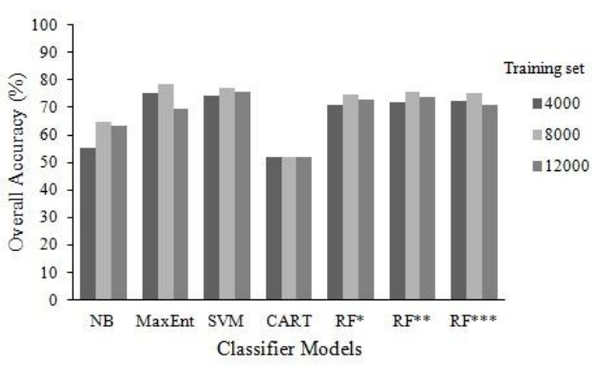

Fig. 2. Classifiers accuracy with 4,000, 8,000 and 12,000 training set sizes.

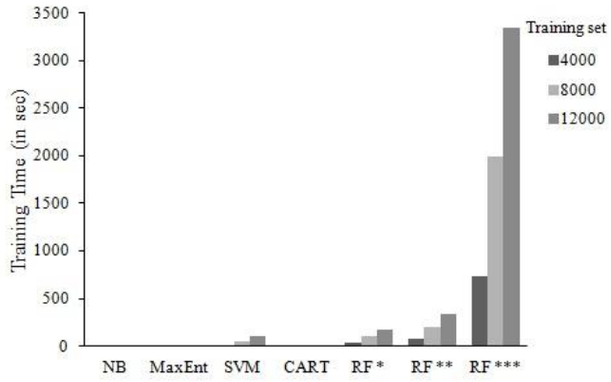

Fig. 3. Training time of each model using the 4,000, 8,000 and 12,000 training set.

Table 1. Real Twitter data classification performance.

Model	Overall Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)	Polarity
MaxEnt	70.5	79.0	81.4	80.2	Positive
		43.8	40.0	41.8	Negative
SVM	73.5	77.8	89.7	83.3	Positive
		50.0	28.6	36.4	Negative
RF	64.4	78.4	71.1	74.6	Positive
		36.3	45.7	40.5	Negative

SUMMARY AND CONCLUSION

CART did not perform well in sentiment analysis in terms of accuracy compared to other learning algorithms but Random Forest (RF) seems to compete very well with other algorithms. RF accuracy is far more superior than NB but relatively lower to MaxEnt and SVM. Overall, the study still presents an interesting result using only a sample training data to train the models. The benefit of this study is to reduce too much dependence on labeled data which is very expensive to acquire and also using only a small sample data for faster, less computational overhead and resource-limited processing of the data. This study would pave the way to consequent studies on sentiment analysis, specifically in the University of the Philippines Mindanao, where no one had yet attempted to explore this new branch of knowledge.

MAIN REFERENCE

Breiman, L., 2001. Random forest [WWW Document]. URL https://www.stat .berkeley.edu/~breiman/randomforest2001.pdf (accessed 11.15.16).

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees. CRC Press, Wadsworth.

Broß, J., 2013. Aspect-Oriented Sentiment Analysis of Customer Reviews Using Distant Supervision Techniques. Freie Universität Berlin.

Go, A., Bhayani, R., Huang, L., 2009. Twitter Sentiment Classification using Distant Supervision [WWW Document]. URL https://www-cs.stanford.edu/people/alecmgo /papers/TwitterDistantSupervision09.pdf (accessed 11.13.15).

Jurka, T.P., Collingwood, L., Boydstun, A.E., Grossman, E., van Atteveldt, W., 2015. RTextTools: Automatic text classification via supervised learning [WWW Document]. Compr. R Arch. Netw. URL https://cran.r-project.org/web/packages /RTextTools/RTextTools.pdf (accessed 12.1.15).

Liu, B., 2012. Sentiment analysis and opinion mining, Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, San Rafael.

Pang, B., Lee, L., 2008. Opinion Mining and Sentiment Analysis. Found. Trends® Inf. Retr. 2, 1–135