Abstract No: 197 **ICT, Mathematics and Statistics**

A REVIEW OF PERFORMANCE OF FEATURE SELECTION METHODS FOR MACHINE LEARNING ALGORITHMS FOR TWITTER SENTIMENT ANALYSIS

M.A.L. Manthrirathna[1*], W.M.H.G.T.C.K. Weerakoon[2] and R.M.K.T. Rathnayaka[2]

[1]Department of Computing and Information Systems, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka
[2]Department of Physical Sciences and Technology, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka
[*]manthrirathna.mal@gmail.com

Twitter sentiment analysis is an area of study with numerous applications. Hybrid approach which combines lexicon and machine learning models are proven to be the best methodology for social media sentiment analysis in previous researches. There are many popular machine learning models suitable for sentiment analysis and their performance varies according to the set of features fed for training. This research aims to review and compare few existing feature selection methods and machine learning models for Twitter sentiment analysis. As the data set for this study 3,378 tweets were collected from Twitter Standard Application Program Interface using the name of a popular mobile phone brand as the search keyword. The initial data set was reduced to 1,709 tweets after the preprocessing step. Then, SentiwordNet lexicon was used to classify tweets as positive, negative and neutral. Features were selected using Recursive feature elimination, Chi-square, Mutual information and F-classification and fed to machine learning models; Multinomial Naïve Bayes, Logistic Regression, K Nearest Neighbors, Decision tree and Random forest. The experiment was repeated 20 times with bootstrap samples to generalize the results. Each sample used for training consisted of 80% of the total tweets and test data set was created using the out of sample tweets. Final results were calculated by averaging the results from all bootstrap samples. The results show that Chi-square, Mutual Information and F-classification methods are accurate and Root Mean Squared Error (RMSE) scores are only slightly different from each other for each machine learning model. Recursive feature elimination shows lower accuracy and higher RMSE score than other methods. Logistic Regression and Multinomial Naïve Bayes have generated the highest accuracies (72.35% and 71.95%) and lowest RMSE scores (0.69 and 0.71). K Nearest Neighbors was the model that generated the lowest accuracy (35.39%) and highest RMSE score (0.85). In conclusion, this study suggests Chi-Square, Mutual Information and F-classification methods are better feature selection techniques than Recursive Feature Elimination for Twitter Sentiment Analysis. Multinomial Naïve Bayes and Logistic Regression were shown to be better classifiers and K Nearest Neighbors was shown to be least suitable classifier for Twitter Sentiment Analysis.

Keywords: Feature selection, Hybrid approach, Machine learning models, Twitter sentiment analysis