A Comparative Study: Best Machine Learning Algorithm for Social Media 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, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka*
[2] *Department of Physical Sciences and Technology, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka*

Sentiment analysis is a field of study that aims to derive the sentiment or the opinion of a text using natural language processing techniques. Performing sentiment analysis on Twitter data has a vast number of applications including predicting stock market prices, product recommendations, etc. Sentiment analysis can be done in lexicon-based, machine learning-based, or hybrid approaches. K Nearest Neighbor, Support Vector Machine, Logistic Regression, Naïve Bayes, K Means Clustering, Decision Trees, and Random Forest are the few most popular machine learning algorithms. This study aims to conduct a comparative analysis among the usage of K Nearest Neighbor, Support Vector Machine, Logistic Regression, and Multinomial Naïve Bayes machine learning algorithms combined with sentword net lexicon to suggest which one provides the best accuracy in sentiment classification of Twitter data. A data set of 1028 tweets was acquired using the Twitter Standard Search API (Application Programming Interface) and Tweepy python library. The name of a popular brand of mobile phones was used to search for tweets. 570 tweets remained after the duplication removal and cleaning process. Then the remaining data was classified as positive, negative, or neutral using sentiword net lexicon and used to train selected machine learning algorithms.80% of the data was used for training and 20% was used for testing. Word counts in the tweets were used as features. Multinomial Naïve Bayes is proved to be the best machine learning algorithm with a model accuracy of 74.56% and K Nearest Neighbor (k=3) is the worst-performing algorithm with an accuracy of 54.38%. Logistic Regression and Support Vector Machine (linear kernel) respectively had accuracies: 72.80% and 70.17%. The result of this research proves Multinomial Naïve Bayes performs relatively better in Twitter sentiment analysis than K Nearest Neighbor, Support Vector Machine, Logistic Regression. This is because two basic assumptions for applying the Multinomial Naïve Bayes algorithm: feature independency and multinomial distribution are well satisfied by the features selected for this study. Also, Multinomial Naïve Bayes can perform well with high dimensional data like tweet text. On the other hand, the poor performance of the K Nearest Neighbor is due to the same reason. K Nearest Neighbor cannot handle a large number of features very well.

Keywords: Sentiment analysis, Twitter, Hybrid approach, Machine learning algorithms, Comparative analysis.