COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS ALONG WITH LEXICAL ANALYZERS FOR SENTIMENT ANALYSIS IN TAMIL LANGUAGE

Shafana, A.R.F.[1], Nihla, M.I.F.[2], Musfira, A.F.[3], Naja, M.M.F.[4]

[1,2,3]Department of Information and Communication Technology, South Eastern University of Sri Lanka, Sri Lanka
[4]Department of Software Engineering, Universiti Malaya, Malaysia

arfshafana@seu.ac.lk

Abstract

The proliferation of social media enables the public to express their views and perceptions readily online. Twitter is one such platform that helps in obtaining a huge amount of textual data and performing useful analysis. Sentiment Classification is one such analysis undertaken to gain insights into public opinion on a certain topic. Although this has been prevalently done using many approaches, the limitations still exist in non-English languages. This study aims to compare the use of the lexical-based approach and machine learning-based approach for classifying the Tamil tweets based on their sentiment. Twitter API was used to perform twitter scraping that resulted in 45852 tweets in total. 300 random tweets were then classified to their respective sentiments by subject experts in the field, this annotated data was used as ground truth and 06 underlying studies were performed on the processed and cleaned data. Four machine learning algorithms (Support Vector Machine, eXtreme Gradient Boosting, Random Forest, and Gaussian Naïve Bayes) and two lexical-based analyzers (VADER and TextBlob) were used for this comparative analysis. The results suggested that the machine learning algorithms performed extremely well where the Support Vector Machine secured the best performance score of all. This study serves as empirical evidence for those interested in performing sentiment analysis on Tamil language tweets.

Keywords: *Machine Learning, Sentiment Analysis, Lexicon, Twitter, Supervised Learning*

ISBN: 978-624-5736-37-9