An Improved kNN Algorithm using K-means and fastText to Predict Sentiments Expressed in Tamil Texts

Authors: Sajeetha Thavareesan[1], Sinnathamby Mahesan[2]

[1]Department of Mathematics, Eastern University, Sri Lanka
[2]Department of Computer Science, University of Jaffna, Sri Lanka

Abstract

With the intention to develop a suitable approach to performing Sentiment Analysis on Tamil Texts using K-means clustering with k-Nearest Neighbour (k-NN) classifier, a corpus UJ_Corpus_Opinions consisting of 1518 Positive and 1173 Negative comments has been constructed. For training 820 positive and 820 negative comments are taken, and for testing 650 and 350 respectively. Bag of Words (BoW) and fastText vectors are used to create feature vectors. These feature vectors are clustered using K-means clustering. The cluster centroids are used as classification keys for k-NN classifier. Two types of clustering techniques are utilised to develop two models: (i) using class-wise information, (ii) with no class-wise information. These two models are tested using K-Fold. All these four models are tested with the two types of feature vectors.

These models are tested using varying number of centroids (Kc:1..10), neighbours (Kn:1..Kc) and folds (Kf:1..10) to study their influence in the accuracy. The accuracy increases with the values of Kc, and the highest accuracy (74%) is obtained for Kn=1 and Kf=2. Accuracy, in general, is found to be more with fastText than with the BoW. The model with fastText and class-wise clustering with K-Fold that obtained 74% accuracy has F1-Score of 0.74.

Key words

Sentiment Analysis, Tamil, K-means, k-Nearest Neighbour and fastText.