Sentiment Analysis on Tamil Articles Using Machine Learning Approaches

Sajeetha Thavareesan

ABSTRACT

Sentiment Analysis (SA) is an application of Natural Language Processing (NLP) to predict the sentiments expressed in a text. In this research work, five different models are experimented to perform SA and identified ensemble based SA model performs better than the other models. Three Tamil corpora and four sentiment lexicons are created along with a list of stop words to experiment proposed models. A novel sentiment lexicon expansion method is proposed to create lexicons of positive and negative sentiment words using *fastText* and *Word2vec* word embedding techniques. A novel word embedding based POS tagger using linear SVM is also proposed along with a method using *Word2vec* and *fastText* word vectors to handle unseen words while performing POS tagging. POS tagging using BoW and TF-IDF with linear SVM give 99% in accuracy for context words with known POS tags and 96% in accuracy for context words with unknown POS tags. Lexicon based SA model is considered as our base model which is obtained 47% and 60% accuracies for the created corpus of opinions and the corpus of movie reviews respectively. The performance of this model is further enhanced by applying preprocessing on texts and by integrating expanded lexicons. Supervised learning based SA model is tested with five different vocabulary construction techniques, among them corpus based vocabulary construction gives impressive performance for *fastText*, *GloVe* and *Word2vec*. In k-means clustering with k-nearest neighbor based approach, four types of experiments are conducted using BoW and *fastText*, wherein *fastText* and class-wise clustering with k-folds of training sets of increased number of clusters can be used to classify the sentiments expressed in the Tamil text. A set of grammar rules along with a sentiment lexicon are used to build grammar rule based SA model. This rule-based model is developed using conjunctions, negational terms and lexicons of positive sentiment words and negative sentiment words and gives a satisfying results. Further an ensemble method using voting technique is proposed which performs better than the other models proposed in this research. Six models are integrated to build this ensemble method. This proposed ensemble method achieved the accuracies of 96% and 78% for movie reviews and opinions corpora respectively. *fastText* is identified as the best feature representation techniques for Tamil. Moreover an efficient POS tagger using Linear SVM and a novel lexicon expansion method are also proposed to enhance the performance of supervised machine learning based SA model along with the construction of corpora and lexicons.

Keywords: Sentiment Analysis, Tamil, Lexicon, Part-of-Speech, fastText.

Recommended by the Supervisor
Dr. S. Mahesan

13 November 2021

S. Mahesan.