Lexicon - pointed hybrid N-gram Features Extraction Model (LeNFEM) for Sentence Level Sentiment Analysis.

James Mutinda[1], Ronald Mwangi[1], and George Okeyo[1]

[1]Affiliation not available

September 18, 2020

Abstract

Sentiment analysis of social media posts and texts can provide information and knowledge that is applicable in social settings, business intelligence, evaluation of citizens' opinions in governance and mood triggered devices in Internet of Things. Feature extraction and selection is a key determinant of accuracy and computational cost of machine learning models for such analysis. Most feature extraction and selection techniques utilize bag of words such as N-grams and frequency-based algorithms especially Term Frequency-Inverse document frequency (TF-IDF). However, these approaches suffer shortcomings such as; they do not consider relationships between words, they ignore words' characteristics and they suffer high feature dimensionality. In this paper we propose and evaluate an approach that utilizes a fixed hybrid N-gram window for feature extraction and Minimum Redundancy Maximum Relevance feature selection for sentence level sentiment analysis. The approach improves the existing feature extraction techniques specifically the N-gram by generating a tri-gram vector from words, Part of speech tags and word semantic orientation. The N-gram vector is extracted by employing a static 3-gram window identified by a lexicon where a sentiment word appears in a sentence. A blend of the words, POS tags and the sentiment orientations of the 3N-gram are used to build the feature vector. The optimal features from the vector are then selected using Minimum Redundancy Maximum Relevance (MR2) algorithm. Experiments were carried out with a publicly available yelp tweets dataset to evaluate the performance of four supervised machine learning classifiers (Naïve Bayes, K-Nearest Neighbor, Decision Tree and Support Vector Machines) when augmented with the proposed model. The results showed that the proposed model had the highest accuracy (86.85%), recall (86.85%) and precision (86.96%).

Hosted file

`Ngram manuscript.docx` available at https://authorea.com/users/360195/articles/481885-lexicon-pointed-hybrid-n-gram-features-extraction-model-lenfem-for-sentence-level-sentiment-analysis