Abstract

Purpose: This is an attempt to classify the level of noise in twitter texts which is part of social media data analytics problem. Estimations in recent machine learning & data feeding algorithms researches’ assumptions consider high data quality in social media texts, while they actually lack data accuracy, completeness, and overall quality which leads to the principle of “Garbage In Garbage Out” resulting in bizarre statistical findings. The aim of this project is to predict and classify Twitter data noise levels using a labelled dataset.

Methodology: After data cleaning, a clustering technique was used to find the major dimensions in the data imported, and a dimension reduction algorithm was ran using PCA Weighting and the Wight Guided Feature Selection algorithms. They resulted into 6 most significant features which were used in the implementation. An artificial neural network model was trained to predict the Tweets’ quality classes using R and RStudio. The ANN used is Neural Network (NN) and Naïve Bayes (NB) for the purpose of predicting the Twitter text quality. There will be a comparison between the 2 ANN used in terms of accuracy and precision.
Findings: Three different aspects of text mining were discovered in twitter data. (1) Neural network gives surprisingly good result as compared to Naive Bayes algorithm, (2) With only 3 hidden layers, a network was created which can predict good or bad class, (3) Preprocessing of the data and implementing predictive algorithms take huge data and very high computational complexity and time. Research results show that Neural Network performs well even without Dropout layer and convolutional layers. The accuracy of the Neural Network is 99%.

References


Index Terms

Computer Science

Artificial Intelligence
Keywords

Data Mining, Twitter Text Quality, Twitter Data Classification, Classification Algorithms, Neural Network Algorithm, Text Analysis.