World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:14, No:12, 2020

Developing an Advanced Algorithm Capable of Classifying News, Articles and Other Textual Documents Using Text Mining Techniques

Authors : R. B. Knudsen, O. T. Rasmussen, R. A. Alphinas

Abstract : The reason for conducting this research is to develop an algorithm that is capable of classifying news articles from the automobile industry, according to the competitive actions that they entail, with the use of Text Mining (TM) methods. It is needed to test how to properly preprocess the data for this research by preparing pipelines which fits each algorithm the best. The pipelines are tested along with nine different classification algorithms in the realm of regression, support vector machines, and neural networks. Preliminary testing for identifying the optimal pipelines and algorithms resulted in the selection of two algorithms with two different pipelines. The two algorithms are Logistic Regression (LR) and Artificial Neural Network (ANN). These algorithms are optimized further, where several parameters of each algorithm are tested. The best result is achieved with the ANN. The final model yields an accuracy of 0.79, a precision of 0.80, a recall of 0.78, and an F1 score of 0.76. By removing three of the classes that created noise, the final algorithm is capable of reaching an accuracy of 94%.

Keywords : Artificial Neural network, Competitive dynamics, Logistic Regression, Text classification, Text mining
Conference Title : ICODAI 2020 : International Conference on Open Data and Artificial Intelligence
Conference Location : Barcelona, Spain
Conference Dates : December 17-18, 2020