World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:16, No:12, 2022

Web Data Scraping Technology Using Term Frequency Inverse Document Frequency to Enhance the Big Data Quality on Sentiment Analysis

Authors : Sangita Pokhrel, Nalinda Somasiri, Rebecca Jeyavadhanam, Swathi Ganesan

Abstract : Tourism is a booming industry with huge future potential for global wealth and employment. There are countless data generated over social media sites every day, creating numerous opportunities to bring more insights to decision-makers. The integration of Big Data Technology into the tourism industry will allow companies to conclude where their customers have been and what they like. This information can then be used by businesses, such as those in charge of managing visitor centers or hotels, etc., and the tourist can get a clear idea of places before visiting. The technical perspective of natural language is processed by analysing the sentiment features of online reviews from tourists, and we then supply an enhanced long short-term memory (LSTM) framework for sentiment feature extraction of travel reviews. We have constructed a web review database using a crawler and web scraping technique for experimental validation to evaluate the effectiveness of our methodology. The text form of sentences was first classified through Vader and Roberta model to get the polarity of the reviews. In this paper, we have conducted study methods for feature extraction, such as Count Vectorization and TFIDF Vectorization, and implemented Convolutional Neural Network (CNN) classifier algorithm for the sentiment analysis to decide the tourist's attitude towards the destinations is positive, negative, or simply neutral based on the review text that they posted online. The results demonstrated that from the CNN algorithm, after pre-processing and cleaning the dataset, we received an accuracy of 96.12% for the positive and negative sentiment analysis.

Keywords : counter vectorization, convolutional neural network, crawler, data technology, long short-term memory, web scraping, sentiment analysis

Conference Title : ICDSBDA 2022 : International Conference on Data Science and Big Data Analytics

Conference Location : Bangkok, Thailand

Conference Dates : December 20-21, 2022