ABSTRACT

Big Data Analysis blended with computational algorithms is a novel tendency in feature abstraction. This involves acquiring knowledge from reliable data sources, rapidity in processing information, and future prediction. Big Data analytics is dynamically evolving with variant features of velocity (analysis time has drastically decreased subsequently), volume (corpus size raise from Big Data to Bigger Data) and Vectors (consonance to dissonance). Organizations now focus on analyzing data that are getting accumulated and are interested in deploying analytics to withstand forthcoming challenges.

Spark is an open source framework for processing contaminated analytics on Big Data. This unified framework gives a wide-range of practices on diverse text data, graph data and structured either static or real time streaming as well. Spark uses MLlib for developing Machine Learning algorithms. These algorithms use less memory, less processing time and largely hand tuned specialized architecture to parallelize large cluster of machines for data analytics.

Machine learning algorithms such as Linear Regression, Decision Tree, Random Forest and Gradient Boosting Tree algorithms are used for analyzing the data sets. The prediction model proposed in this research, is applied to interpret the data sets by using the machine learning algorithms and to analyze the best forecast value from the comparative study.

The main objective of this research work is to find the best prediction from the machine learning techniques using the model proposed. Various approaches have implemented the supervised and unsupervised techniques using MapReduce approach but the proposed model uses Apache Spark framework for comparing the existing methods. This approach computes the best prediction from the model by evaluating the time complexities with each technique. This thesis focuses on highlighting the features of datasets for analyzing the best prediction through machine learning algorithms.