ISSN 2319 – 7595
Volume 8, No.2, March - April 2019
International Journal of Information Systems and Computer Sciences
Available Online at http://warse.org/IJISCS/static/pdf/file/ijiscs03822019.pdf
https://doi.org/10.30534/ijiscs/2019/03822019

Intelligent Analysis methods on BigData: A Comparison

Sharon Susan Jacob[1], **Dr.R.Vijayakumar**[2]
[1] Research Scholar, School of Computer Sciences,
Mahatma Gandhi University,
Kottayam, Kerala, India
sharonsusanjacob@gmail.com

[2]Professor, School of Computer Sciences,
Mahatma Gandhi University,
Kottayam, Kerala, India
vijayakumar@mgu.ac.in

ABSTRACT

Big Data systems are the eventual outcomes of the today's data centric World. Big Data is becoming more popular with the internet technology development, which means the data sets whose size is beyond the ability of current technology, method and theory to capture, manage, and process the data within a tolerable elapsed time. A major source of Big data during the last decade originated from Internet related applications such asSocial media. Twitter is one of the largest social media site, because it is having millions of tweets every day. Twitter tweets are some of the central source of unstructured data and to find meaningful information from unstructured data is very difficult. With the help of classification technique, it is possible to change unstructured data into organised form. In this work, twitter data was taken as the dataset. For training data, Intelligent analysis methods like Random Forest, Logistic Regression and Decision Tree classifier are used. For analysing the data, Python programming was used. The result shows that Random Forest classifieryielded more classification accuracy than Logistic Regression classifier andDecision Tree classifier.

Key words: BigData, Decision Tree classifier, Logistic Regression,Python,Random Forest.

1. INTRODUCTION

The recent years has witnessed data flow from a multitude of sources and this led to the Big Data revolution. Big Data [1,2]starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data. Big data enables organizations to store, manage, and manipulate vast amounts of data at the right speed and at the right time to gain the right insights. Therefore, big data is the capability to manage a huge volume of disparate data, at the right speed, and within the right time frame to allow real-time analysis and reaction.There are three main types ofdata that makeup Bigdata: structured, semi-structuredand unstructured data. Big data is broken down by three characteristics, viz., **Volume, Velocity and Variety.** Volume refers to the amount of Data that is getting generated. Velocity points out the speed at which data is getting generated. Variety indicates the different types of data that is getting generated. Fig 1. shows Characteristics of Big Data.

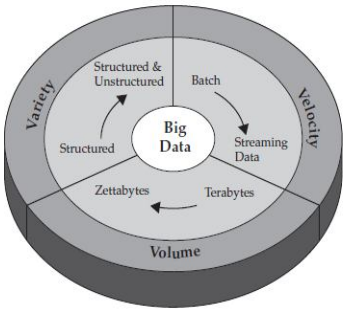

Figure 1: Characteristics of Big Data

As data has become the fuel of growth and innovation, it is more important than even to have an underlying architecture to support growing requirements. But before delve in to the architecture, the data must first be captured[3,4] and then be organized and integrated. After the successful implementation of this phase, data can be analyzed based on the problem being addressed. Finally, management takes action based on the outcome of that analysis. Figure 2 shows the cycle of Big data management.

Figure 2: The cycle of Big data management

3. METHODOLOGY

Large amount of data is increasing every day because of social media and Twitter is one of the social media site which gives an opportunity to people to express their ideas and opinions about a particular topic. Twitter tweets and other social media posts are some of the central source of unstructured data. So twitter data [6]is taken as the Data set.The data set after pre-processing and feature extraction, the data is given to machine learning model. The intelligent methods used are Random Forest, Logistic Regression and Decision Tree classifier and the language used is Python. The performance of these algorithms are analysed.

The *pre-processing* of the data is a very important step as it decides the efficiency of the other steps down in line and it involves removal of unimportant features from the data. The steps involved should aim for making the data more machine readable in order to reduce ambiguity in feature extraction. Pre-processing involves removal re-tweets, stemming, lemmatization etc. The quality and quantity of features is very important as they are important for the results generated by the selected model.Selection of useful words from tweets is *feature extraction.*

3.1 Intelligent methods applied for training the data set

Random Forest, Logistic Regression and Decision Tree classifier are the methods implemented for preparing and training the data. ***Random forests*** are probably the most accurate classifiers being used today in machine learning. They can be easily parallelized, making them efficient to run on large data sets, and can handle a large number of features, even with a lot of missing values.***Logistic regression*** is a classification model in which the response variable is categorical. It is a statistical method for analysing [5]a dataset in which there are one or more independent variables that determine an outcome. In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, fraud etc.) or 0 (FALSE, failure, not-fraud, etc.). A ***Decision Tree*** is an algorithm used for supervised learning problems such as classification or regression. A It is a tree in which each internal (nonleaf) node is labelled with an input feature. The arcs coming from a node labelled with a feature are labelled with each of the possible values of the feature. Each leaf of the tree is labelled with a class or a probability distribution over the classes.A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions.

4. ANALYSIS OF THE DATA

Twitter sentiment analysis was done using the language Python. Open source programming such as Python provide high quality implementation of numerous data analysis and visualization method, from regression to statistics, text analysis and much more. It uses white space inundation to delimit blocks and provides a large[7] standard library, which can be used for various applications. In using the Language Python, the packages - SCIKIT-LEARN, NumPy and Pandas are used. The ***Scikit-learn*** project started as scikits. learn, a Google Summer Code project. It is a powerful library that provides many machine learning classification algorithms, efficient tools for data mining and data analysis. The various functions that can be performed by this library are classification, regression, clustering, preprocessing etc.***NumPy*** is the fundamental package for scientific computing with Python. It provides a high-performance multidimensional array object, and tools for working with these arrays. In computer programming, ***panda***sis a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

5. PERFORMANCE EVALUATION

The performance analysis of Random Forest, Logistic Regression and Decision Tree classifier on Big data is done. The methods were employed with the collected samples of datasets containing 1 lakh samples of twitter information. The metrics used here for evaluating the performance are accuracy, precision, recall and F1-score.

$$Precision = \frac{tp}{tp+fp} \quad (1)$$

$$Recall = \frac{tp}{tp+fn} \quad (2)$$

$$F1\ Score = 2 * \frac{(recall * prcision)}{(recall + prcision)} \quad (3)$$

where, t_p is true positive which correctly predicted positive values, t_n is true negative which correctly predicted negative values, f_p is false positive which is falsely predicted positive class and f_n is falsely predicted negative class.The metric representation is demonstrated in the table 1 below:

The accuracy obtained with Random Forest classification technique on Bigdata is 70% whereas Logistic Regression and Decision Tree classifier produces 69% and 66% accuracy respectively. The precision and recall measures obtained by the Random Forest method are 0.71 and 0.70 where as Logistic Regressionproduced 0.69 precision & 0.69 recall values andDecision Tree classifier produced 0.67 precision and 0.67 recall values. The F1 score obtained by the Random Forest method is 0.70 whereas Logistic Regressionand Decision Tree classifier produces 0.69 and .67 scores. Thus it is evident from the table that Random Forest Classifier yielded more classification accuracy than Logistic Regression classifier and Decision Tree classifier.

Table 1: Metric Representation of Big Data (Twitter data)

Technique	Accuracy	Precision	Recall	F1 score
Random Forest	70%	.71	.70	.70
Logistic Regression	69%	.69	.69	.69
Decision Tree Classifier	66%	.67	.67	.67

6. CONCLUSION

Big data mining is an emerging trend in the field of data science. It is the arising need for various engineering domains.Twitter sentiment analysis comes under the category of text and opinion mining. It focuses on analyzing the sentiments of the tweets and feeding the data to a machine learning model in order to train it and then check its accuracy. It comprises of steps like data collection, text pre-processing, sentiment detection, sentiment classification, training and testing the model. The classification accuracies of Random Forest, Logistic Regression and Decision Tree classifier on the twitter data is compared and the result shows that Random Forest yielded more classification accuracy than Logistic Regression classifier and Decision Tree classifier.

REFERENCES

[1] Breiman L, **Random Forests**, Machine Learning, 45, 5-32, (2001).https://doi.org/10.1023/A:1010933404324

[2] Ahmed Fuad Mohammed, Dr. Vikas T. Humbe, Dr.Santosh S. Chowhan,"**A Review of Big Data Environment and Its Related Technologies"**, International Conference On Information Communication And Embedded System(ICICES 2016).
https://doi.org/10.1109/ICICES.2016.7518904

[3] M.Trupthi, Suresh Pabboju, G.Narasimha, **"Sentiment Analysis On Twitter Using Streaming API",**2017 IEEE 7th International Advance Computing Conference, January 2017, pp. 915-919.
https://doi.org/10.1109/IACC.2017.0186

[4] Jiawei Han, MichelineKamber and Jian Pei, **"DATA MINING Concepts and Techniques"**, Morgan Kaufman Publishers, USA, (2002).

[5] Grover, Purva and Johari, Rahul.**BCD: BigData, cloud computing and distributed computing. Global Conference on Communication Technologies**, GCCT 2015. doi-10.1109/GCCT.2015.7342768.

[6] Geetika Gautam, Divakar Yadav, **"Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis"**, 2014 Seventh International Conference on Contemporary Computing (IC3), August 2014, pg.437-442.
https://doi.org/10.1109/IC3.2014.6897213

[7] A. Machanavajjhala and J.P. Reiter, **"Big Privacy: Protecting Confidentiality in Big Data,"**ACMCrossroads, vol.19, no. 1, pp. 20-23, 2012.
https://doi.org/10.1007/s15016-012-0046-2