# Chapter 51
# A Novel Approach to Optimize the Performance of Hadoop Frameworks for Sentiment Analysis

**Guru Prasad**
*SDMIT, Ujire, India*

**Prithviraj Jain**
*SDMIT, Ujire, India*

**Amith K. Jain**
*SDMIT, Ujire, India*

**Nagesh H. R.**
*A.J. Institute of Engineering and Technology, Mangalore, India*

## ABSTRACT

*Twitter is one among most popular micro blogging services with millions of active users. It is a hub of massive collection of data arriving from various sources. In Twitter, users most often express their views, opinions, thoughts, emotions or feelings about a particular topic, product or service, of their interest, choice or concern. This makes twitter a hub of gargantuan amount of data, and at the same time a useful platform in getting to know and understand the underlying sentiment behind a particular product or for that matter anything expressed in twitter as tweets. It is important to note here that aforesaid massive collection of data is not just any redundant data, but one which contains useful information as noted earlier. In view of aforesaid context, Sentiment analysis in relation to twitter data gains enormous importance. Sentiment analysis offers itself as a good approach in classifying the opinions formulated by individuals (tweeters) into different sentiments such as, positive, negative, or neutral. Implementing Sentiment analysis algorithms using conventional tools leads to high computation time, and thus are less effective. Hence, there is a need for state-of-the-art tools and techniques to be developed for sentiment analysis making it the need of the hour to facilitate faster computation. An Apache Hadoop framework is one such option that supports distributed data computing and has been commonly adopted for a variety of use-cases. In this article, the author identifies factors affecting the performance of sentiment analysis algorithms based on Hadoop framework and proposes an approach for optimizing the performance of sentiment analysis. The experimental results depict the potential of the proposed approach.*

DOI: 10.4018/978-1-6684-6303-1.ch051

## 1. INTRODUCTION

In today's digital world social networking sites play a vital role and also have an influential say in modern way of life. Twitter is one among the most popular social networking sites with more than 100 million of daily active users. According to Statista survey, as of year 2017 Twitter had 328 million active users and the number is said to have increased and still increasing day by day (Andreas et al.,2017). In Twitter, registered users can read and post tweets; tweets are limited to 280 characters. They can also upload images and short videos of size not more than 5MB and 512MB respectively. Millions of users express their views, opinions, thoughts, emotions, feelings about different products, events, people, etc., on the twitter platform.

Indian Premier League (IPL) is a popular, professional Twenty-Twenty (T20) cricket league played in India. It ranks sixth among all sports leagues across the world. As we already know cricket in India is not just viewed as a sport, but, a religion in itself. Due to its humungous popularity, unending reach along with an uncanny ability to arouse interest and then being able to follow it up with definite action, it evokes all sorts of emotions, feelings and what not among cricket viewers. The same goes true for IPL, its fans, and in general, viewers of IPL. In Twitter, IPL fans originating from various places express their views, opinions, thoughts, emotions or feelings about their favorite IPL teams and players. During IPL season millions of tweets get tweeted every day on a regular basis. Aforesaid live stream of data is considered to be a rich source of information for Sentiment analysis. Natural Language processing is used to mine people's opinions about IPL teams and players expressed in form of tweets. Sentiment analysis helps in classifying people's opinions as positive, negative or neutral Implementing Sentiment analysis algorithms using traditional data analytics tools seem unable to handle Twitter Big Data as data to be handled is humongous, changing at a fast pace and characteristically complex by nature. Big data analytics has modernized traditional data analytics by introducing new technologies that support distributed storage and processing of large amount of data. Today, Apache Hadoop has become a highly popular and powerful distributed computing framework to process large amounts of data. It is composed of Hadoop Distributed File System (HDFS), Yet Another Resource Negotiator (YARN) and MapReduce parallel programming model. The unique features of Hadoop that make it so attractive are ease of access, robustness, fault tolerance, scalability and ease of parallel programming. Using Hadoop framework, a lot of work has already been proposed on Sentiment analysis in relation to Twitter data. However, some parameters affecting the performance of Sentiment analysis remain a challenge on Hadoop framework. When working with large amounts of data sets, there will be challenges and difficulties such as data sets consuming more HDFS disk space, network related issues and high computation time. In this paper, the author identifies the factors affecting the performance of sentiment analysis algorithm based on Hadoop framework and proposes an approach for optimizing the performance of sentiment analysis. Experimental results obtained show that proposed novel approach effectively optimizes the HDFS disk space utilization, speeds up the data movement in the network and optimizes the computation time.

The rest of the paper is organized as follows: Section 2 comprises of literature survey in relation to the proposed topic; Section 3 presents the proposed framework and associated implementation so as to optimize the performance of sentiment analysis with regard to Twitter data; Section 4 substantiates aforesaid analysis by showcasing comprehensive experimental results; Finally, Section 5 delivers conclusion to the paper.