

A Comparative Study of Sentiment Analysis of Social Media Streams Using Big Data Technologies

Sharma Arushi Premchand¹, Syed Yawar Ahmed², Hasan Phudinawala³

P.G. Student, Department of Data Science¹⁻²

Coordinator, Department of Data Science³

Royal College of Arts, Science and Commerce (Autonomous), Mira Road (East)

Abstract: *The rapid growth of social media data has made sentiment analysis a central part of digital intelligence rather than just a computational problem. With billions of daily opinions being generated by users, scalable and real-time analytics have been in demand, which has led to the use of many Big Data technologies, including Apache Hadoop, Apache Spark, and Apache Flink. The given paper aims at providing a non-technical comparison of these three frameworks in the social media sentiment analysis context. It has used thirteen peer-reviewed research publications and six real-world case-studies to compare each platform to each of seven important parameters: architecture and processing model, speed and latency in data processing, ease of use, machine-learning and NLP support, scalability and fault tolerance, best use case, benefits, drawbacks and cost implications. The analysis indicates that Hadoop is still dependable on massive-scale batch analytics, whereas Spark offers a balanced ecosystem of iterative machine-learning piping and nearly real-time operations and Flink can be used to provide better stream processing with event-time semantics and low latency. These benefits are, however, interdependent to the complexity of operations and cost of resources. The paper finds that the best framework is based on the focus of analytics: Hadoop to aggregate sentiment on a large scale offline, Spark to perform hybrid analytics and train models, and Flink to monitor social-media in real time and detect events.*

Keywords: Sentiment Analysis, Social Media Analytics, Apache Hadoop, Apache Spark, Apache Flink, Big Data Technologies, Stream Processing

