Chapter 69
Real-Time Streaming Data Analysis Using a Three-Way Classification Method for Sentimental Analysis

Srinidhi Hiriyannaiah
Department of Computer Science and Engineering, Ramaiah Institute of Technology, Bangalore, India

G.M. Siddesh
Department of ISE, Ramaiah Institute of Technology, Bangalore, India

K.G. Srinivasa
Department of Information Technology, Ch Brahm Prakash Government Engineering College, New Delhi, India

ABSTRACT

This article describes how recent advances in computing have led to an increase in the generation of data in fields such as social media, medical, power and others. With the rapid increase in internet users, social media has given power for sentiment analysis or opinion mining. It is a highly challenging task for storing, querying and analyzing such types of data. This article aims at providing a solution to store, query and analyze streaming data using Apache Kafka as the platform and twitter data as an example for analysis. A three-way classification method is proposed for sentimental analysis of twitter data that combines both the approaches for knowledge-based and machine-learning using three stages namely emotion classification, word classification and sentiment classification. The hybrid three-way classification approach was evaluated using a sample of five query strings on twitter and compared with existing emotion classifier, polarity classifier and Naïve Bayes classifier for sentimental analysis. The accuracy of the results of the proposed approach is superior when compared to existing approaches.

DOI: 10.4018/978-1-7998-2460-2.ch069

1. INTRODUCTION

In the advent of the big data, characterized with 3V's namely velocity, variety and volume different forms of data exist. They include structured, semi-structured and unstructured data. A new form of data that is emerging in this paradigm is streaming data. The examples of streaming data are data from sensors, twitter feeds, weather data, air quality data, etc. Streaming data are often produced in real-time (Arasu et al., 2016) (Krempl et al., 2014). The processing of streaming data is quite challenging due to the rate at which the data are generated in the form of streams. The processing of other data forms such as textual/ numerical involves physical simulation and computations are normally disk based (Arasu et al., 2016). Such type of processing techniques cannot be used on streaming data. For example, consider a data stream with 3 datanums, calculation on the datanum2 has to be completed before going to datanum3.A system that performs streaming data analytics should focus on distributed approach for storing and analysis of streams.

The streaming elements are often generated at an extreme rapid rate, which needs to process in real time without accessing the archival storage. Conventional approach of usage of relational database systems (RDBMS) necessitates deposit of objects with insertion, deletion and updating turning up less reiteration than queries (Hashem et al., 2015). The other issues related to RDBMS have prompted research to augment existing technologies and build new systems to manage data streams.Thus, there is a need for a framework and system that handles the different issues related streaming data effectively. The proposed system addresses these different issues related to streaming data analytics.

Social computing is the analysis and modelling of activities that takes place on various social networking platforms. It is one of the innovative and exemplar growing research area in the field of computing. Intellectual and interactive applications are developed to produce and derive efficient results (King et al., 2009). Due to the wide availability of social network sites, people share their opinions and views about an event, product or issue. An insight into such informal and unstructured data is highly useful to draw some conclusions in various fields such as disaster relief and humanitarian assistance, marketing and trade predictions, checking political polls, advertising market, scientific surveys, checking customer loyalty, finding job opportunities, population health care and understanding students' learning experiences (Wang, 2007; Kambatla et al., 2014; Dredze, 2012).

In social computing, sentimental analysis (SA) is an ongoing field of research in computer science that deals with the study of human behaviour towards a real-world entity in terms of opinions. It is used to solve many data mining problems related to classification (Medhat et al., 2014). Social networking sites such as Twitter, Facebook, LinkedIn, etc., are used by most of the people in today's interconnected world. Information from these sites can be used to carry out SA. Twitter is a popular social networking site that has become more widely used among users and varieties of topics are discussed. It is used as an endorsement for organizations to increase their brand awareness and reputation (Fan et al., 2014) Twitter feeds commonly called as tweets are unstructured in nature. It does not allow more than 140 characters, thus users use vague way of expressing their views such as "Hey @xyz !! Great Win!!". In order to perform analytics over such kind of data entity recognition, parsing of data are needed for insights.

Sentimental analysis is used in various fields for understanding consumer behaviour and mainly used for recommendations (Chen et al., 2014; Medhat et al., 2014; Liu, 2012). One of the approaches used for sentimental analysis is knowledge based approach that is based on lexicon partitioning techniques. In this type of approach, lexicons are used for extracting the opinions and determine the polarity of the sentiment (Bahrainian et al., 2013; Gautam et al., 2014; Maks & Vossen, 2012). With machine learning