Sentiment Analysis of Social Media Data using Hadoop Framework: A Survey

Anisha P. Rodrigues
NMAM Institute of technology
Nitte,India

Niranjan N. Chiplunkar,
PhD
NMAM Institute of technology
Nitte,India

Anujna Rao
NMAM Institute of technology
Nitte,India

ABSTRACT
Nowadays, online social media have become the important platform across the globe to share information. People prefer online social media as it is easy to share their opinions on a daily basis hence sentiment analysis is of utmost importance wherein people rely on the opinions shared online. With this extensive growth in the usage of online social media, huge amount of social data is generated. How to process this large set of data efficiently, effectively and in a manner suitable for the user is an important research topic. In this paper, we firstly introduce the definition of sentiment analysis as well as Hadoop and describe the Hadoop architecture, then focus on the analysis of Hadoop framework for sentiment analysis of social media data.

General Terms
Hadoop, Hadoop Distributed File System, MapReduce, Opinion, Sentiment analysis, Social media data.

Keywords
Apache Pig, Hive, Sqoop, HBase, Zookeeper, Flume

1. INTRODUCTION
With the extensive growth in the usage of online social media, the ransom amount of data is available as users' preference regarding any product, services provided by various organizations or with respect to any political issues. Micro blogs, forums are also available wherein internet users, can express their opinions. Since mobile devices can access network easily from anywhere, social media is becoming more and more popular. The number of people using the social media is increasing day by day as they can share their personal feeling every day and reviews are created in large-scale. Every minute opinion, reviews are being expressed online and a potential user rely on these reviews, opinions, feedback given by various other users to make decisions with respect to purchasing an item or developing a software when it comes to an organization that provides services. Analyzing these reviews, opinions or feedback in this scenario is of utmost importance. It seems evaluating these reviews, opinions are not as easy as it appears to be, and it requires performing sentiment analysis. Sentiment analysis greatly helps us in knowing the customer behavior. The biggest challenge is to process the social data which are in unstructured or semi-structured form. The former technologies fail to process the data in this form in an effective way. So, there is a need for highly optimized, scalable and efficient technology to process the abundant data that are being produced at a high rate. The social media data produced will be either unstructured or semi-structured. Hadoop framework effectively analyzes the unstructured and semi-structured form data.

1.1 Sentiment analysis
Sentiment is defined as an expression or opinion by an author about any object or any aspect. Analyzing, investigating, extracting users' opinion, sentiment and preferences from the subjective text is known as sentiment analysis. The main focus of sentiment analysis is parsing the text. In simple terms, sentiment analysis can be defined as detecting the polarity of the text. Polarity can be positive, negative or neutral. It is also referred to as opinion mining as it derives opinion of the user. Opinions vary from user to user and sentiment analysis greatly helps in understanding users' perspective. Sentiment can be,

Direct opinion as the name suggests the opinion about an object is given directly and the opinion may be either positive or negative. For example, "The video clarity of the cellphone is poor" expresses a direct opinion.

Comparison opinion, it is a comparative statement which consists of comparison between two identical objects. The statement, "The picture quality of camera-x is better than that of camera-y" is one possible example for expressing a comparative opinion.

Sentiment analysis is performed at three different levels:

- Sentiment analysis at sentence level identifies whether the given sentence is subjective or objective. Analysis at sentence level assumes that the sentence contains only one opinion.
- Sentiment analysis at document level classifies the opinion about the particular entity. Entire document contains opinion about the single object and from the single opinion holder.
- Sentiment analysis at feature level extracts the feature of a particular object from the reviews and determines whether the stated opinion is positive or negative. The extracted features are then grouped and their summarized report is produced.

1.2 Apache Hadoop
For processing the large sets of data in parallel across cluster of nodes Apache came up with an open source framework known as Hadoop. The major components of Hadoop are Hadoop distributed file system (HDFS) and the MapReduce programming model. Hadoop is accessible, because it runs on cloud computing services or commodity machine across clusters of nodes. It is able to handle failures in an efficient manner even though it is intended to run on commodity hardware which makes it robust. Any number of nodes can be added to the Hadoop cluster in order to deal with huge data in parallel. Hadoop is simple in that a user can write a simple parallel code. To each and every node data is distributed and hence operation is performed in parallel in Hadoop cluster. Hadoop overcomes the hardware failure by keeping multiple

copies of data. The following sections of the paper contains the Hadoop framework architecture in the section II, related work on sentiment analysis of various social media data using Hadoop in the section III, and finally conclusion in section IV.

2. HADOOP FRAMEWORK ARCHITECTURE

Modules of Hadoop are as follows [11]:

- Hadoop Common utilities: Hadoop modules require operating system level and file system level abstractions which are provided by the java libraries and utilities. Execution of Hadoop is carried out by the java files and scripts facilitated by Hadoop common utilities.
- Hadoop Yarn framework: Scheduling of jobs and managing the resources to the clusters is carried out by the Hadoop Yarn.
- Hadoop Distributed File System (HDFS): File system of Hadoop stores very huge set of data and provides easier access to this stored data, resulting in high throughput.
- MapReduce paradigm: MapReduce enables parallel processing of the data.

The two main components of Hadoop are MapReduce and HDFS are discussed below.

2.1 MapReduce

The MapReduce paradigm enables the writing of applications in an effective manner and also processing of huge sets of data is efficient with Hadoop MapReduce [11]. The MapReduce paradigm has two different tasks:

- The Map Task: The Map task captures the input and this input data are divided into pair of data. This data is further divided into tuples to form a key/value pair.
- The Reduce Task: The input to the Reduce task is the output from the Map task. All the divided tuples in the Map task is combined to form smaller set of tuples. Map Task is followed by Reduce Task.

The MapReduce component of the Hadoop framework schedules monitors the tasks and also re-executes the failed task. The MapReduce paradigm has a single JobTracker and one TaskTracker that acts as master and slave respectively. The master JobTracker directs the slave TaskTracker to execute the task and also it manages the resource, tracks the resource distribution, consumption and availability. On the other hand the TaskTracker provides the status information to the JobTracker.

2.2 Hadoop Distributed file system

Hadoop provides its own filesystem known as Hadoop distributed file system based on Google File Server (GFS) which is highly fault-tolerant. The architecture of HDFS depicts the master/slave architecture. Master node manages the file system and the storage of actual data is taken care by the slave node.

A file in a HDFS namespace is divided into several segments and these segments are stored in DataNodes. The plotting of these segments to the DataNodes is identified by the NameNode. The data node performs read and write operations.

3. HADOOP FRAMEWORK FOR SENTIMENT ANALYSIS

Hadoop proves to be a reliable framework and also it processes the huge set of data in a fault-tolerant manner which makes it efficient. Many companies face the difficulty in obtaining the customer feedback about their products, considering this scenario, in paper [1] authors have proposed a computational framework for fast feedback opinion mining. The real-time Twitter data stream is given as input to the framework to filter and analyze the obtained data in order to provide quick feedback through sentiment analysis. For sentiment analysis data accuracy is very important and Apache Hadoop framework provides an accurate result of 84% when data produced from social media are abundant. The services are provided by Cloudera version of Hadoop. With the help of Cloudera manager, services like Flume, Hive and Oozie which were installed on top of Hadoop. A demerit of this paper was that, the sentiment dictionary used for text analysis did not give them accurate result. Instead words were divided into nine categories which possibly gave them the efficient answers. The authors of paper [2] Songtao Shang, Minyong Shi, Wenqian Shang, Zhiguo Hong have proposed a system that processes the public opinion using mahout algorithms that possibly executes in Hadoop framework. Here TF-Gini algorithm is used for pre-processing and for processing Hadoop mahout is used. Mahout, a data mining algorithm that uses MapReduce paradigm, mainly designed to process bulk of data based on Hadoop.

In paper [3] an illustration on the usage of open source technologies for sentiment analysis from Facebook data is done by the authors. A significant advancement in the in-memory computation capabilities can be obtained through Spark with Resilient Distributed Datasets in the open source world. The combination of Spark RDD and Hadoop provides significant computational capabilities in a fault tolerant cluster setup with low price commodity hardware. This platform provides an information analytics layer on top of Hadoop that embraces the MapReduce paradigm and the resilience of Spark RDD's along with advanced statistical analysis layer through R with design time and run time optimizations of the open source stack. Although this paper provides an approach that mines unstructured sentiment information, refinement in this approach can be made through the iterative process of cleaning /pre-processing, which would ultimately eliminate the outliers and noise from the data source.

In paper [4], authors have proposed a faster retrieval approach of sentiment analysis wherein Hadoop is used to store and process the large set of data. When Hadoop is implemented with bloom filter it is possible to retrieve the results of sentiment analysis at a faster rate and also in an accurate manner. The application with bloom filter can grow at high rate and we can analyze the sentiment easily, providing accurate results. Bloom filter is a data structure supporting add, find and sometimes delete operation widely used for testing if element is in a set, especially if the set is huge. Authors of paper [5] addressed the possible challenges faced when Apache Hadoop is implemented as a Big data platform for opinion mining approach. In this paper, a design of cooperative architecture is described and is implemented between Hadoop and social media in order to effectively deal with the huge data and also the issue of collecting and storing of data. Also, from written Arabic language users' preferences are analyzed using an efficient technique that covers various kinds of user thoughts. A case study is discussed that covers the most interesting topic in Middle East region (MERS-CoV

infection) to evaluate the proposed methodology [5]. While designing they faced a challenge in constructing the polarity lexicon as it was the main component in analyzing, classifying sentiments. Data source used for this approach is twitter data stream. The responsibility of moving the data to HDFS and also aggregating the collected data is effectively handled by Flume. The proposed approach in the paper can be easily applied to different domain.

Much research is being done in the construction of recommendation system which facilitates user in understanding users' preferences and also in finding favorable, new subject based on the pattern produced by users' ratings and ranks, in paper [6] authors discuss a recommendation system, that provides a summary of users' reviews, comments, feedback about any subject using Hadoop framework. Recommendation describes ratings and ranks for various items which are in numeric term. Mahout Interface is implemented to analyze the data in the form of ratings and ranks. Author of this paper predict that possible summary on the reviews could be provided based on the ratings and ranks given by the user. The work in the paper [7], describes the performance analysis of apache Hive for query execution of Twitter tweets in order to calculate MapReduce CPU time spent and total time taken to finish the job. Authors of this paper have also made a comparative study of Big data technologies along with their features, Technologies such as Apache Pig, Hive, Sqoop, HBase, Zookeeper, Flume, are integrated with Hadoop to increase the efficiency and performance of Hadoop. The experiments done by this paper shows that, when the number of nodes is increased the MapReduce CPU time spent increases but there is a decrease in Hive query execution time which is an advantage. So, they concluded saying that all the technologies implemented on top of Hadoop improves the performance of basic Hadoop MapReduce Framework. Future work of this paper is the implementation of the technologies that they mentioned to improve the performance of Hadoop MapReduce Framework.

In paper [8] authors have developed a single Big data platform for social T.V. analysis that extracts the important views from T.V. social response in a real-time manner. Usually we face challenges while extracting the real-time data especially in networking architecture and it would be very effective if the network is reconfigurable and also if virtual machines are provided on demand. So the authors of this paper have described a cloud-centered platform with software designed network (SDN). Mainly it has three components, robust data crawler system, an SDN enabled Big data processing system and a social media analytics system [8]. Integrating SDN with Hadoop increases the processing rate. The described solution is built on Hadoop and Storm platform for real-time analysis. Hadoop integrates HDFS, HBase and MapReduce. The inability of Hadoop in supporting cross-site shuffle is overcome with the combination of SDN controller and Hadoop Job scheduler that enables flow-forwarding feature. To deal with problem of processing the huge set of data, authors of paper [9] aim at designing a framework for deriving and analyzing the climate of opinion about an item from social network. They have mentioned various frameworks to process the large sets of data but they concluded that Hadoop is the perfect one to deal with huge sets of data. They have provided an overview of framework architecture. MapReduce paradigm is used which is most methodical and presents an error forbearance mechanism. The data ordering consists of Hadoop/MapReduce, HBase as database and Apache Flume to collect the data from the social media. The framework designed here extracts and analyzes the opinions for social customer relationship management (CRM). The framework proposed here gives general opinion about the client fulfillment, but this can further be ameliorated by providing opinion about each property of the product. To dynamically recommend services to the users a review is done by authors in the paper [10] based on recommendation system. To make this recommendation system more effective and robust Hadoop framework is used. In this paper top-k services recommendation list is provided by performing sentiment analysis on the rating values. Here, sentiment analysis is done by extracting keywords from passive users' reviews and a rating value is given to every new keyword in the dataset. The proposed recommendation system is based on user-based collaborative filtering algorithm, key-words extracted from passive users' review which is used to indicate user preferences. Porter-stemmer algorithm is used to extract the keyword in root form. Sentimental analysis is used to calculate total rating based on this rating top-k services are recommended to the users. They have given the general concept of recommendation system method and also the proposed system architecture. In order to increase the scalability and efficiency the proposed recommendation system is built on top of Hadoop.

4. DISCUSSION and FUTURE WORK

In this paper, the existing work done on sentiment analysis using Hadoop framework have been discussed .With the increasing dependence on social media data, the information obtained from the web in the form of feedbacks, comments have gained much attention in the field of sentiment analysis. The major challenge with this extensive growth in the usage of social media is processing and analyzing the huge sets of data produced as a result. With the implementation of MapReduce paradigm, Hadoop framework proves to be a reliable framework as it processes the huge sets of data in a fault- tolerant manner. We can implement the technologies such as Apache Pig, Hive, Sqoop, HBase, Zookeeper, and Flume on top of Hadoop in-order to improve the efficiency and performance of Hadoop. The combination of Spark RDD and Hadoop also improves the in-memory computation capabilities.

5. REFERENCES

[1] Lokyamanyathilak Govindan Sankar Selvan and Teng-Sheng Moh, "A framework for fast-feedback opinion mining on twitter data streams." Collaboration Technologies and Systems (CTS)(2015):314-318.

[2] Songtao Shang, Minyong Shi, Wenqian Shang, Zhiguo Hong,"Research on public opinion based on Big Data." Computer and Information Science(ICIS)(2015):559-562.

[3] Sudipto Shankar Dasgupta, Swaminathan Natarajan, Kiran Kumar Kaipa, Suja Kumar Bhattacherjee and Arun Viswanathan, "Sentiment analysis of facebook data using hadoop based on open source technologies." Data Science and Advanced Analytics(DSAA)(2015):1-3.

[4] Devendra K Tayal and Sumit Kumar Yadav, "Fast retrieval approach of sentiment analysis using bloom filter hadoop." Computational Techniques in Information and Communication Technologies(ICCTICT)(2016):14-18.

[5] Anis Zarrad, Abdulaziz Aljaloud and Izzat Alsmadi, "The evaluation of the public opinion." Utility and Cloud Computing(UCC)(2014):664-670.

[6] Jai Prakash Verma,Bankim Patel, Atul Patel "Big Data Analysis: Recommendation System with Hadoop Framework."Computational Intelligence and Communication Technology(CICT)(2015):92-97.

[7] Aditya Bhardwaj, Vanraj, Ankit Kumar, Yogendra Narayan, Pawan Kumar, "Big Data Emerging Technologies: A CaseStudy with Analyzing Twitter Data using Apache Hive."Recent Advances in Engineering and Computational Sciences(RAECS)(2015):1-6.

[8] Han Hu, Yonggang Wen, Yue Gao, Tat-Seng Chua, and Xuelong Li "Toward an SDN-Enabled Big Data Platform for Social TV Analytics," *IEEE Network.*,vol.29,pp. 43-49, Sept-Oct. 2015.

[9] Fatima Zohra ENNAJI, Abdelaziz EL FAZZIKl ,Mohamed SADGAL, DjamalBENSLIMANE, "Social Intelligence Framework: Extracting and Analyzing Opinions for Social CRM." Computer Systems and Applications(AICCSA)(2015):1-7.

[10] Khushboo R. Shrote, Prof. A.V.Deorankar,"Review Based Service Recommendation for Big Data".Advances in Electrical, Electronics,Information, Communication and Bio-Informatics(2016):470-474.

[11] http://www.tutorialspoint.com/hadoop/hadoop_introduction.htm

.