Analyzing social media with InfoSphereBigSheets

Marouane Birjali[1], Abderrahim Beni-Hssane[1],and MohammedErritali[2]
[1]LAROSERI Laboratory, Department of Computer Sciences, University of ChouaibDoukkali, Faculty of Sciences, El Jadida, Morocco
[2]TIAD Laboratory, Department of Computer Sciences, University of Sultan MoulaySlimane, Faculty of Sciences and Technologies, BéniMellal, Morocco

Abstract:*Nowadays the term Bigdata becomes the buzzword in every organization due to ever-growing generation of data every day in life. Big data is high in volume, velocity and also high in variety that is, it can be a structured, semi-structured, or unstructured data. Big data can reveal the issues hidden by data that is too costly to process and perform the analytics such as user's transactions, social and geographical data issues faced by the industry. In this paper, we investigate twitter, which is the largest social networking area where data is increasing at high rates every day is considered as big data. This data is processed and analyzed using InfoSphereBigInsightstool, which bring the power of Hadoopin real time. This also includes the visualizations of analyzing big data charts using BigSheets.*

Keywords:*Social media, Twitter Data; Big Data; Hadoop; Infosphere; BigSheets;*

1. Introduction

The companies today face growing challenges from their commercial perspective in particular their like to value should be produced from huge amount of data generated and on the complexity of the data that is both structured and semi-structured and unstructured. During the last years, the Internet has yet seen a wider scope through the development of social media. Based on easy communication techniques and accessible to all, the media promote social interaction through the Internet. Offering free access, social media has greatly promoted the mass and have triggered public debate on the Internet. Many social networks exist and there are more than 900 social media sites available on the internet [1]. Millions of people are using Twitter and it is ranked as one of the most visited sites with the average of 58 million tweets per day [2]. Big data is the border of the ability of an enterprise to store, process and access all the data it needs for the effective functioning, make decisions, reduce risks, and serve [3] customers within a time reasonable time.

The first organizations to embracing Big data were online and startup businesses. Companies like Google, Twitter, eBay, LinkedIn and Facebook were erected around Big Data from the beginning. Big data can achieve reductions dramatic cost, substantial improvements in the time necessary to effect a spreadsheet task [4]. According to the statistical in the real time industries, there are 2.5 million items added per minute by each individual. Similarly 300,000 tweets, 200 million e-mails, pictures generate 220,000 per minute, and other companies as 5TB of RFID and is larger than 1PB data for gas turbines producing daily. In 2012 all of these data is 2.8 zettabytes only now to know in 2015, it raised to 20 zettabytes and by 2020, this number may reach 40 zettabytes. In this world, 90% of unstructured data and is becoming difficult to treat broadcast on for business. the organizations and individual companies should understand more deeply overcoming this problem. The data is continues to rise, is moving too quickly or does not come within the structures of database architectures. For earning the value of big data, another approach is needed to treat [5].

This massive data is considered as a Big data and can be used for industrial or business purpose after organizing as per the need and processing. This paper addresses how to analyze social media data using BigSheets for streamed, processed, analyzed and visualized using Apache Hadoop and other tools in a distributed fashion. Of the many Big Data technologies, Hadoop is popular and widely used in meeting the Big Data challenges. There are several different platforms providing Hadoop for enterprises to disseminate their data as Apache Hadoop, the BigInsights IBM, Microsoft Azure HD Insights, tools Cloudera and Hortonworks etc. All these tools divers perform functions for analysis of the data based on the different problem domains.

This paper is organized as follows. Some related work on topic model in section 2 and our methodology

of work is presented in section 3. In section 4, the suicidal acts analysis model is described. Section 5 describes experiments setup and the result. Finally, the paper is summarized briefly in section 6.

2.Related Works

In this chapter, we describe the related research and the study of social media analysis. The social network has attracted the attention of the research community that is trying to understand, among others, their structure and user interconnection and interaction between users. People tend to express their feelings and talk about their activities of daily life through Twitter.

There has been a wide range of research done on sentiment analysis, from rule-based, bag-of-words approaches to machine learning techniques. Two main research directions of opinion mining operate on either the document level [6,7,8] or the sentence level [9,10,11]. Both document level and sentence level classification methods are usually based on the identification of opinion words or phrases. For this, there are basically two types of approaches: (1) lexicon-based approaches, and (2) rule-based approaches. In the former, a lexicon table will be built first, with each word in this table belonging to positive or negative evaluations. The echo count of positive words and negative words will then be calculated, or some formula which considers the distance between each opinion word and product feature will be used to determine the sematic direction. In rule-based approaches, parts-of-speech (POS) taggers will first be used to tag each word, and then co-occurrence patterns of words and tags will be found to determine the sentiments.

Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. There has been a wide range of research done on sentiment analysis [12], from rule-based, bag-of-words approaches to machine learning techniques. Starting from the document level classification task Turney [13] it has been handled at the sentence level Hu and Liu [14] and more recently at the phrase level Wilson et al. [15].

The social network like Twitter, on which users post his reactions to and opinions about "everything", is a new and different challenge. Some of the first results and recent analysis of Twitter sentiment data. Two main research areas of mining opinion operate either on the document level [16]. Both classification methods at the document level and at the level of the sentence are generally based on the identification of opinion words or phrases.

However in this paper, we focus on social media data, on which users post real time reactions to and activities about “everything”, where this data is increasing at high rates every day is considered as big data. This data is processed and analyzed using InfoSphereBigInsightstool which bring the power of Hadoop to the enterprise in real time. This also includes the visualizations of analyzing big data charts using big sheets and workbooks.

3. System architecture

3.1. Apache Hadoop

Apache Hadoop (High-availability distributed object oriented platform) is a distributed system that addresses these issues. On the one hand, it offers a storage sys-tem via its HDFS distributed file system (Hadoop Distributed File System) and it offers the possibility of storing the data in the duplicating, a Hadoop cluster therefore does not need to be configured with a RAID system becomes useless. On the other hand, provides a Hadoop data analysis system called MapReduce. It plies on the HDFS file system to perform processing on large data volumes [17].

The MapReduce architecture is comprised of two phases of phases- map and re-duce phase. Initially, monitoring of the use of input data can be divided into several copies as <key, value> where the key is the word and the value indicates how many of times the word has occurred and assigns to each underemployment the task track-ers. Next, mix the values arrange in an order. Finally, in the phase cut, the results of each job Tracker are combined to produce the final results [18].

HDFS is highly fault tolerant, which is designed using low-cost hardware holds up very large amount of data is stored on multiple machines (Multi-nodes). Some of the important characteristics of HDFS is the storage and processing in a distributed environment, streaming access to data files and providing the security system through file permissions and authentication. The following figure shows the architectural over-view of HDFS and Mapreduce.

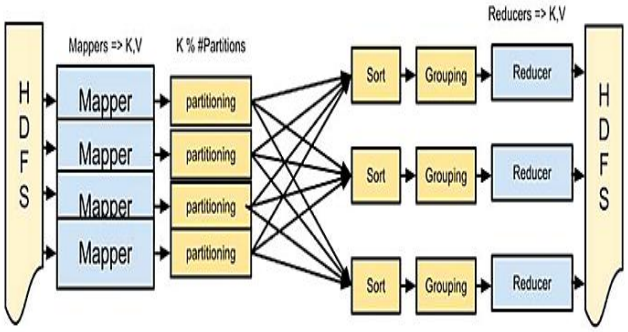

Figure 1. MapReduce and HDFS interaction.

3.2.IBM's Platform - InfosphereBigInsights Architecture

IBM InfoSphereBigInsights provides massive volumes analysis functions of data on one platform for businesses. It combines the open source Apache Hadoop solution with enterprise features and integration, and provides a large-scale analysis, characterized by its resilience and fault tolerance. The software supports structured data, unstructured and semi-structured in their native format, providing maximum flexibility. It is provides advanced analytics based on Hadoop technology to meet the requirements analysis of massive data volumes. And Designed for performance and ease of use with performance optimized functions, visualization, rich development tools and powerful analytical functions. Integrates with IBM and other information solutions to simplify and improve data manipulation tasks [19] [20].

The Figure 2 illustrates the IBM's Big data platform, which integrated software libraries for processing, streaming, and persistent data.

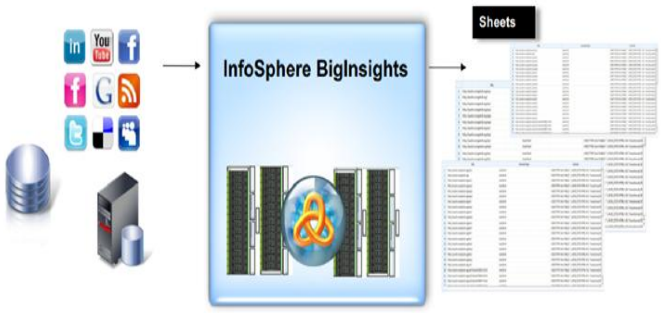

Figure 2. IBM's Big Data Platform Architecture.

4. Problem Statement and Methodology

4.1.Existing tools

As mentioned above, small set of social media data can be downloaded and easily treated through traditional databases. Procedure using ancient techniques stream and analyze the raw data. Also uses these ancient logical coding to filter and find positive words, negative and moderate the text collection of files and stores them in databases. However, sometimes the data can be huge in quantity and unstructured raw data which traditional databases cannot handle, process and analyze. This is the major problem comes when disseminating and processing large volumes of data in real-time enterprises. This problem could be addressed by constructing a dashboard to super-vise the sentiment of Twitter traffic related to a given topic in near-real time, allowing users to enjoy near real-time Twitter sentiment of business ideas or any other purposes using Apache Hadoop ecosystem.

This paper discusses how to overcome the limits of traditional techniques utilizing Hadoop ecosystem to streamline the processing of data from large clusters. To disseminate and analyze easy data of twitter, Apache Hadoop ecosystem and IBM BigInsights tools could be used. Twitter data is analyzed from the API twitter, processed using Jaql script and stored in HDFS and analyzed, negative words as positive using methods for MapReduce and finally it returns Tweet ids result then displays the results as graphics using BigSheets tool BigInsights.

4.2.Strategy for Word Sense Disambiguation

In this paper, we will resolve the question of big data using Hadoop and ecosystem platform in order to simplify the data processing. Here we have used Apache Hadoop and IBM BigInsights platform to collect and analyze Big Data problems easily. Here are a social media Big Data sample such as twitter reaches the level of treatment finally showed the results that different arrays using bigsheets tool BigInsights.

For analyzing twitter, the proposed system involves the following steps in order to overcome the problem encountered in the existing system:

- *Creating Twitter applications*: the step of play and retrieve the twitter data is to create an application using the twitter API by logging into Twitter account to Retrieved the Consumer Key (API Key), Consumer Secret (API Secret), Access Token and Access Token Secret.
- *Tweets collection data by using Twitter4j:*is a Java library for Twitter API. It allows to easily integrate a Java application with the twitter service. We have used it to collect tweets using its streaming and search methods implementation from the twitter4j package [21].
- *Processing data with Jaql and Mapreduce:*The data collected is unstructured format. Jaql the script is used to extract important data,

transforming them into a simpler structure to convert the delimited file by commas, and then storing the data in HDFS to perform the processing using MapReduce.

- *Creating workbook and graphics master BigSheets:*For the graphical visualizations and final data manipulations, we use BigSheets. BigSheets is a spreadsheet-style tool provided with BigInsights that allows you to use standard spreadsheets functions, join tables, filter data, sort data, and visualize data in graphs.

The Figure 3 above shows total percentage coverage by languages. To achieve this result, we just basically group data from the Tweets Data and provide the total count of tweets in every group.

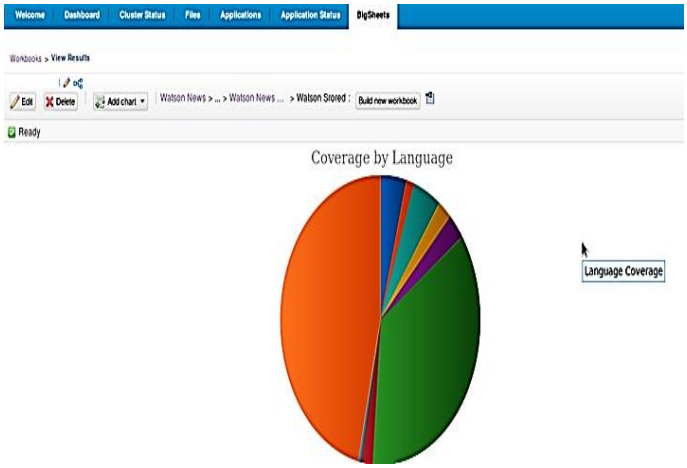

Figure 3. Coverage by language in a pie chart.

The Figure 4 present the BigSheets Twitter Analysis, number of tweets during the time, and for the figure 5 depicts a tag cloud chart we generated for the words. As with any BigSheets tag cloud, larger font

indicates more occurrences of the data value and scrolling over a data value reveals the number of times it occurred in the collection.

Figure 4.Visualization by BigSheets Twitter Analysis for Number of tweets during the time.

Figure 4. Visualization by BigSheets Twitter Analysis for Tag cloud.

5. Conclusions and Future Work

As part of this work, we present a way of streaming, analyzing and visualizing the twitter data using BigInsights system and also the integration of it with Apache Hadoop. The above said tools not only applicable for streaming, processing, analyzing, and visualizing the twitter data but also could be enhanced to apply other types of Big data from various sources. In this paper, it is concluded that processing time for analysis of massive twitter data is minimized and retrieving capabilities are also made efficient by using the proposed method when compared to other traditional processing methods for Big data.

References

[1] Statistic Brain. Twitter Statistics, Retrieved from http://www.statisticbrain.com/ twitter-statistics/, 2014.

[2] R. Li, K. H. Lei,R. Khadiwala, Chang, "TEDAS: A Twitter-based Event Detection and Analysis System," icde, pp.1273-1276, 2012 IEEE 28th International Conference on Data Engineering, 2012.

[3] Peter Lake, Paul Crowther, "Concise Guide to Databases: A Practical Introduction", Springer-Verlag London 2013.

[4] Thomas H. Davenport, "Big Data at Work: Dispelling the Myths, Uncovering the Opportunities", Harvard Business Review Press, Boston, Massachusets.

[5] O'Reilly Radar Team, Planning for Big data, A CIO's Handbook to changing the Data Landscape
[6] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86, 2002.
[7] P. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. ACL'02, 2002.
[8] K. Dave, S. Lawrence, and D. Pennock. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. WWW'03, 2003.
[9] M. Gamon, A. Aue, S. Corston-Oliver, and E. K. Ringger. Pulse: Mining customer opinions from free text. IDA'2005.
[10] M. Hu and B. Liu. Mining and summarizing customer reviews. KDD'04, 2004.
[11] S. Kim and E. Hovy. Determining the Sentiment of Opinions. COLING'04, 2004.
[12] G. Vinodhini and R. Chandrasekaran, "Sentiment analysis and opinion mining: A survey," International Journal, vol. 2, no. 6, 2012.
[13] P. Turney, "Thumbs up or thumbs down Semantic orientation applied to unsupervised classification of reviews," Proceedings of the Association for Computational Linguistics.
[14] Hu M. and Liu B., "Mining and Summarizing Customer Reviews", KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 168-177.
[15] Wilson T., Wiebe J. and Hoffmann P., "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis", In the Advanced Research and Development Activity (ARDA).
[16] Wu, Yuanbin, Qi Zhang, Xuanjing Huang, and Lide Wu. Structural opinion mining for graph-based sentiment representation. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP-2011). 2011.
[17] http://hadoop.apache.org/
[18] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Google, Inc.
[19] Steven Hurley, James C. Wang, IBM System x Reference Architecture for Hadoop: IBM BigInsights Reference Architecture.
[20] http://www.ibm.com/developerworks/data/library/techarticle/dm-1110biginsightsintro/
[21] www.twitter4j.org 12 http://git-scm.com/