ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

Big-SoSA:Social Sentiment Analysis and Data Visualization on Big Data

Mahalakshmi R[1], Suseela S[2]

PG Scholar, Department of CSE, Periyar Maniammai University, Tamil Nadu, India[1]

Assistant Professor, Department of CSE, Periyar Maniammai University, Tamil Nadu, India[2]

Abstract: In recent years, due to the popularity of social networking has dramatically increased and the vast amount of data being produced by social networks such as Twitter, Facebook, Google+, etc.,. Social networks become popular among millions of people who share's their thoughts in everyday life. Social media web sites are rich source of data for sentiment analysis. Sentiment Analysis, has been used to understand the people's opinion on particular product or service. Twitter, one of the biggest and most popular social web site which contains unstructured data. In order to analysis such a data we need effective methodology which can process huge volume of data. Therefore, in this paper, we propose a method of sentiment analysis on twitter by using Hadoop and its ecosystems that will process the large volume of data on a Hadoop and the MapReduce function will perform the sentiment analysis.

Keywords: Social media; Sentiment analysis; TwitterAPI; Hadoop; MapReduce; Flume;

I. INTRODUCTION

In every day, the social networks generate massive amount of data that have led to the data explosion. Big Data is considered as very large volume of data sets which can be found easily on web, social media, remote sensing data, medical records and industrial etc. A basic task in sentiment analysis is classifying the polarity of a given text or document [20]. Most of the social network data's are unstructured and semi-structured data. Social networks generate vast amount of data and the data size is huge, analysing such amount of data is difficult. Social networks data can be used for some other purpose like marketing, understanding customer behaviour, etc., .Since the data format is relatively easy and free, most social network data are unstructured. Unstructured data may be defined as data that has not been standardized, because its structure and shape are so complex, unlike video image data and document data [17].

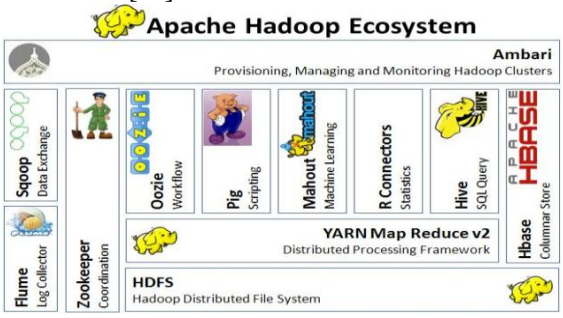

Figure 1 Hadoop Achitecture

In order to analysis large amounts of unstructured data on a social sites, we need effective processing technology. There are so many open source platform available for this big social data processing. The Hadoop ecosystem is a famous big data processing system. Hadoop is an open source distributed software platform for storing and processing data. Hadoop has rapidly emerged the standard for managing large volumes of unstructured data [23].the following figure 1 shows the Hadoop Architecture and its Ecosystems. There by we introduced SoSA method (Social sentiment analysis) it is used to extract meaningful information from social networks. This system used Map Reduce function to identify the polarity and score.

The rest of the paper organized as follows. In Section II a short overview of related works is given. In Section III we introduced SoSA system architecture is given. Then in Section IV describes Sentiment analysis procedure. Subsequently in Section V, given the experiment and results. Paper is conclude in Section VI.

II. RELATED WORKS

Most of the social network data's are unstructured and semi-structured data. To analyses such a data sets are difficult in real time analysis. Park et.al [1] explained how to automatically collect a twitter corpus for sentiment analysis and opinion mining. Using the corpus, they build sentiment classifier that is able to determine positive, negative and neutral sentiments for a document level analysis. They contributed three important concepts first, collect the corpus which is consist of both positive and negative. Second, the described how to perform statistical linguistic analysis of those collected corpus. Third, build the sentiment classifier [10]. Hao et al [24]. Described the approach that attempts to automatically analyses large volumes of twitter comments with respect to what was commented on positively or negatively. Using natural language processing techniques, to determine topics, extract attributes of the topics, detect opinions about the attributes, and measure the sentiment value. Sunil B et al [6] explained that they concentrated more on the speed of performing analysis than its accuracy. They performed the emoticons and hashtags for the sentiment analysis. They used emoticons but the use of hashtags to determine the context of the tweet is not done.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

III. PROPOSED SYSTEM

HDFS and MapReduce: To handle a variety of unstructured social data efficiently, a big data processing system is proposed. HDFS, which is based on the Hadoop ecosystem, is used to reliably collect and store data from a large amount of social data. MapReduce is used to effectively analyse large amounts of unstructured data as to the sentiment of the user. HDFS is a file processing system that has a distributed processing structure. MapReduce is a software framework developed by Google to support distributed computing. In this paper, we use map function to perform polarity pre-processing analysis. The algorithm for polarity pre-processing is shown in Algorithm 1.The lower part of Figure 2 shows the proposed MapReduce functions and the processes for sentiment analysis. Twitter data collected through the online streaming tool called flume which is based on Hadoop ecosystem. Then the collected data loaded into HDFS for storage and MongoDB used to save the status of twitter data. Sentiment analysis performs a polarity pre-processing function. This examines the context in each sentence to enhance the accuracy and subjects them to pattern matching with the negative context dictionary and the positive context dictionary. Sentiment analysis by using MapReduce data dictionary is composed of polarity and score. Finally, we used R language for visualizing the sentiment score and polarity of twitter data. For coding part we used Java, Python (text classification using nltk package) and R for data visualization. Figure 2 shows the proposed architecture for sentiment analysis. In following session explained how to get twitter raw data by using TwitterAPI and how to get real time streaming data by using flume online streaming tool.

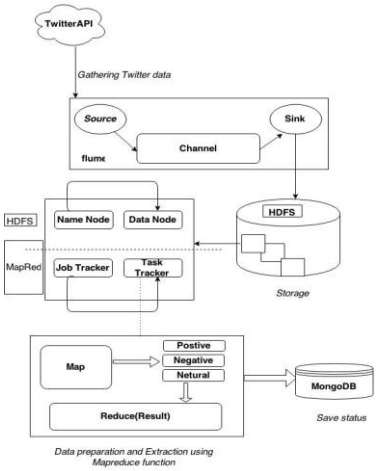

Figure 2: Proposed SoSA System Architecture

IV. METHODOLOGY

In this section, the SoSA proposed methodology is explained. The following for steps are used to analysis twitter data on Hadoop. First, the data is gathered from twitter website by using TwitterAPI. Second, the data extraction process will take place. Third, the extracted data is processed to load into the HDFS storage. Fifth, sentiment analysis is perform by using the MapReduce function.

A. Data Collection

Data set is collected by using TwitterAPI and flume. Flume is online streaming tool which is used to move or load large volume of data. We need flume configuration to define what kind of data that we want. in order to configure flume we need to set source as a TwitterAPI ,sink as HDFS. Figure 2 shows the structure of flume. Flume gather the data from twitter by using secrets keys which is provided by twitter application developer site. Figure 3 shows the processing of TwitterAPI.

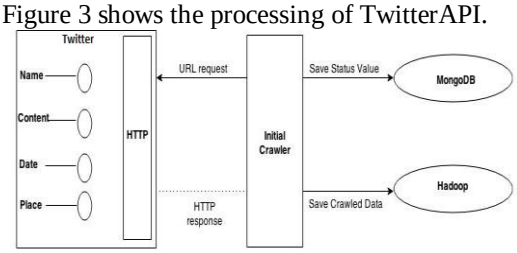

Figure 3: Data gathering using TwitterAPI

B. Feature Extraction

The data collected from Twitter contain a lot of unnecessary data such the stop word, unstructured data (hapyyyy to happy), hashtag. Therefore, from collected data we need only the necessary data by removing stop word, hashtag. Table 1 shows feature extraction of raw data as an example.

Extraction Data	Content
Text	Tweets
Name	User name
Time	Creation time of tweets

Table 1: Example of Data Extraction

C. Polarity Check and Score.

The extracted data is used analysis MapReduce function to identify the polarity of twitter data .Figure 2 show the MapReduce function .The MapReduce function consist of Map and Reduce. The map function used to cross check the positive context dictionary and negative context dictionary. Finally the reducer function will have mapper result and this result is used to save twitter status into mongo DB.

V. EXPERIMENT AND RESULT

In our experimental analysis of twitter data is processed to check performance of the proposed methodology. The following tests were carried out i,e Polarity test and sentiment score which result's sentiment analysis of the given data set . The following figures shows the result which obtained by our experimental system, with the use of R language we visualized the result. Figure 4 shows the twitter volume separated by positive tweets and negative tweets and the polarity score.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

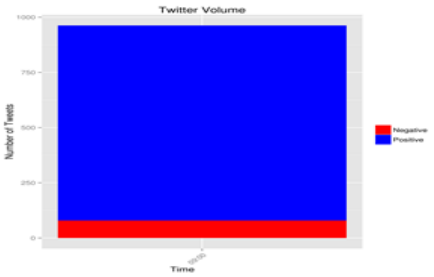

Figure 3

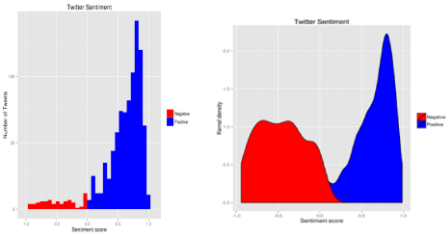

Figure 4

Finally we have the twitter status which is saved in MongoDB. Its shows the Number of tweets and its polarity .the following table 2 shows the Twitter status. Figure 5 shows the polarity negative score as well as positive score. Table 3 shows the average scoring for both positive and negative tweets.

Number (%) of positive tweets	89.02%
Number (%) of negative tweets	7.96%
Number (%) of neutral tweets:	3.02%

Table 2

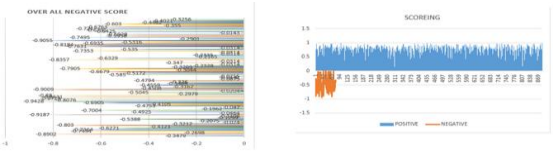

Figure 5

VI. CONCLUSIONS

Social networks nowadays become very popular among internet users and it's generate huge volume of data every day. This type of data can be used for some valuable purpose. In order to analyses social data we need effective methodology which can handle analysis of huge volume of data sets as well as real time analytics. Thus, the proposed SoSA system is able to collect meaningful information from the twitter and effectively perform sentiment analysis .This method is composed of a HDFS system based on the Hadoop ecosystem and MapReduce functions. The MapReduce function is used to classify the twitter sentiment polarity and score. For analysis of social media, data gathered by using TwitterAPI and flume. We used R language for visualizing the result of social sentiment and MongoDB is used to save the twitter status.

REFERENCES

[1]. A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC '10), 2010.
[2]. A.Kowclka, Aditi Gupta, Karthick Sondhi, Nishit Shivhre, Raunaq Kumar "Sentiment analysis for social media" Volume 3, Issue 7,International Journal of Advanced Research in Computer Science and Software Engineering. pp.216-221,2013.
[3]. Ayushi Dalmia ,Mayank Gupta, Arpit Kumar Jaiswal Sunil and Chinthala Tharun Reddy [online] "Sentiment Analysis in twitter" Available at :http://researchweb.iiit.ac.in/.
[4]. B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval,vol.2,no.1-2, pp.1–135,2008.
[5]. Bing Liu, Sentiment Analysis and Opinion Mining, Morgan and Claypool Publishers, May 2012.pp.18-19, 27-28,4445,47,90-101.
[6]. B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde "Real Time Sentiment Analysis of Twitter Data Using Hadoop" (2014) Volume 3, International Journal of Computer Science and Information Technologies, pp-3098 – 3100.
[7]. Data Visualization [online] Available at:http://www.sas.com/en_us/insights/big-data/data-visualization.html.
[8]. Efthymios Kouloumpis , Theresa Wilson, and Johanna Moore, "Twitter sentiment analysis: the god the bad and the OMG!," in Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, pp. 538–541, 2011.
[9]. Flume, http://flume.apache.org/
[10]. G.Vinodhini, RM.Chandrasekaran. "Sentiment analysis and opinion Mining: A Survey " , Volume 2, Issue 6,International Journal of Advanced Research in Computer Science and Software Engineering.2012.
[11]. Hadoop, http://hadoop.apache.org/.
[12]. Ilkyu Ha, Bonghyun Back, and Byoungchul Ahn,"MapReduce functions to analyze sentiment information from social big data" International Journal of Distributed Sensor Networks.Id-417502,2014
[13]. J. Han, E. Haihong, G. Le, and J. Du, "Survey on NoSQL database," in Proceedings of the 6th International Conference on Pervasive Computing and Applications (ICPCA '11), pp. 363–366, October 2011.
[14]. Jalaj S. Modha, Gayatri S. Pandi, Sandip J. Modha "Automatic Sentiment Analysis for Unstructured Data" Volume 3, Issue 12,International Journal of Advanced Research in Computer Science and Software Engineering.2013.
[15]. J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," Communications of the ACM,vol.51,no. 1, pp. 107–113, 2008.
[16]. K. Chodorow, MongoDB: The Definitive Guide, O'REILLY, 2nd edition, 2013.
[17]. McKinsey, Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey & Company, 2011, http: //www.mckinsey.com/.
[18]. M. Bautin, C. B. Ward, A. Patil, and S. S. Skiena, "Access: news and blog analysis for the social sciences," in Proceedings of the19th International World Wide Web Conference (WWW'10),pp.1229–1232, April 2010.
[19]. Stephen Whitworth -Python, R, machine learning and data visualisation "Sentiment anaysis in python using NLTK"(marcg 31-2013) Available at :http://www.sjwhitworth.com/.
[20]. Sentiment analysis, http://en.wikipedia.org/wiki/Sentiment_analysis
[21]. V. N. Khuc, C. Shivade, R. Ramnath, and J. Ramanathan,"Towards building large-scale distributed systems for Twittersentiment analysis," in Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC '12), pp. 459–464, March 2012.
[22]. Walaa Medhat,Ahmed Hassan ,Hoda Korashy "Sentiment analysis algorithms and applications: A survey",Ain shams Engineering Journal.2014.
[23]. White Paper Big Data Analytics Extract, Transform, and Load Big Data with Apache Hadoop-Intel corporation.
[24]. Ming Hao, Christian Rohrdantz, Halldór Janetzko, Umeshwar Dayal Daniel A. Keim, Lars-Erik Haug, Mei-Chun Hsu Hewlett-Packard Labs and University of Konstanz "Visual Sentiment Analysis on Twitter Data Streams"(october-2011),IEEE Symposium on Visual Analytics Science and Technology Providence, Rhode Island, USA 978-1-4673-0013-1/11/©2011 IEEE.