ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

A Sentiment Analysis of Twitter Data using Hadoop Framework

Jeevitha, Raksha, Vipin N

Lecturer, Department of Computer Science, Srinivas School of Engineering, Mukka, Mangalore, Karnataka, India

B.E., Department of Computer Science, Srinivas School of Engineering, Mukka, Mangalore, Karnataka, India

B.E., Department of Computer Science, Srinivas School of Engineering, Mukka, Mangalore, Karnataka, India

ABSTRACT: Twitter is one of the Social Networking Sites which contains enormous useful data. In this paper, how sentiment analysis of twitter data is carried out is discussed. Here we use Hadoop platform to increase the efficiency and scalability of the analysis using a parallel processing procedure called Map-Reduce. Finally, substantial procedures are carried out on large real-world data sets to achieve considerable accuracy than the existing techniques.

KEYWORDS: Twitter, HDFS, Hadoop, Map-Reduce, Sentiment Analysis

I. INTRODUCTION

Twitter is an online news and social networking service where users post and interact with messages. Registered users can post tweets, but those who are unregistered can only read them. Users access Twitter through its website interface, SMS or a mobile device app. It enables users to hand out, modify and rank the content, as well as to convey their personal opinions about specific topics.
In this project, we implement a system in Hadoop which analyses twitter data. Twitter data is in the form of comments which are nothing but sentiments i.e., opinions/feelings of people. This data is collected by using Twitter API. Tweets are normalized and then map-reduce technique is carried out. By analysing this data, our system will give output in the form of positive and negative polarity of tweets. Here, it makes use of data dictionary for classifying the data. And this analysed data can be represented in the form of graphs.

II. RELATED WORK

In [2], authors proposed a system to analyse the sentiments of Twitter users through their tweets in order to extract what they think. They used hadoop for sentiment analysis which will process the huge amount of data on a hadoop cluster faster. In [3],the authors took advantage of large datasets available from Twitter micro blogging platform and widely available stock market records. Data was collected during three months and processed for further analysis. Machine learning was employed to conduct sentiment classification of data coming from social networks in order to estimate future stock prices. Calculations were performed in distributed environment according to Map Reduce programming model. Evaluation and discussion of results of predictions for different time intervals and input datasets proved efficiency of chosen approach were discussed. In [4], the authors did data analysis ontweets about movies to predict several aspects of the movie popularity. The main results they presented were whether a movie would be successful at the box office.

III. PROPOSED ALGORITHM

A. Design Considerations:

- Ram – Minimum 8GB
- Ubuntu Operating system for high efficiency.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

B. *Description of the Proposed Algorithm*

Aim of the proposed system is to build an application which analyses the twitter data and gives the positive and negative polarity of the tweets for a given hashtag as a keyword with greater accuracy and efficiency than the existing system. It includes 5 steps:

Step 1:Retrieval of Data
The twitter data is extracted from the twitter API.
Step 2: Preprocessing
The collected data is stored in the HDFS and then normalized so that the special characters or unwanted data is removed.
Step 3:Tweet Correction
Here the data which contains short-hands are replaced with the exact word.
Step 4: Polarity Detection
Using Map-Reduce technique the data is broken down into separate words with its count. Then it is matched with the data dictionary which contains values for the words to detect how many positive and negative words are present for a particular topic.
Step 5: Emotion Extraction
Finally the data is classified according to the polarity detected in the previous step and the result is displayed in the form of graph.

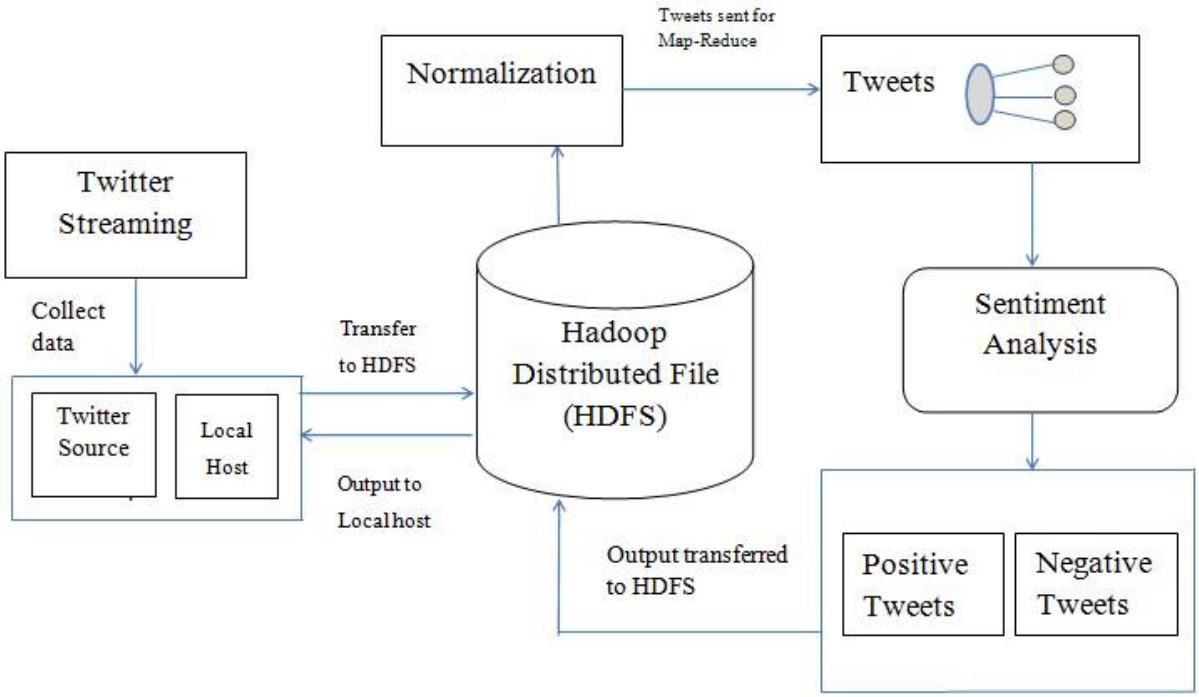

Figure: Architectural Diagram of the System

IV. PSEUDO CODE

Step 1: Start
Step 2: Input a keyword (hashtag).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Step 3: Extracting the tweets for the given keyword.
Step 4: Storing the extracted tweets in the HDFS.
Step 5: Normalizing the tweets (Removal of unwanted characters).
Step 6: Data is subjected to Map-Reduce technique.
Step 7: Sentiment Analysis – Finding the Polarity of the words and matching them with the data dictionary.
Step 8: Classifying the tweets according to the polarity.
Step 9: Displaying the result using graph(pie chart).
Step 10: Stop

V. SIMULATION RESULTS

The simulation involves the sentiment analysis of tweets for a particular keyword and output will be in the form of pie chart with percentage of positive and negative polarities. Figure 1. shows a frame for entering the hashtag which is to be analysed. Figure 2. shows the tweets downloaded for the given hashtag. Figure 3. shows storing of downloaded tweets from local host to HDFS. Figure 4. shows the normalized tweets i.e., removing the special characters, links, at signs, hash symbols etc. Figure 5. shows the final output after map-reduce and sentiment analysis of the normalized tweets in the form of percentages for positive and negative polarities using apie- chart.

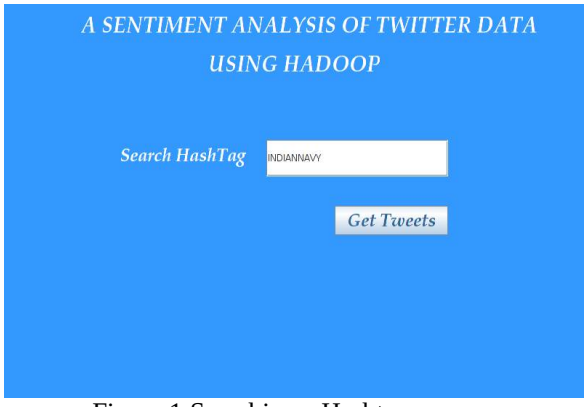

Figure 1:Searching a Hashtag

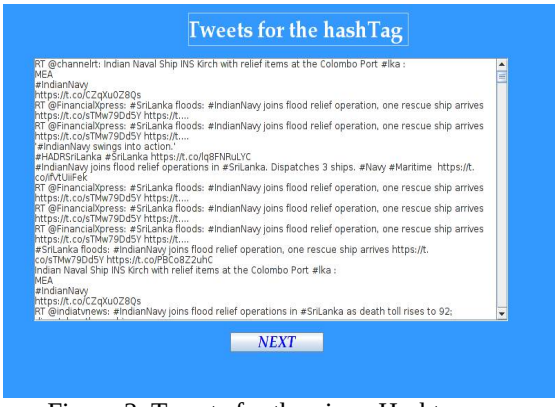

Figure 2: Tweets for the given Hashtag

Figure 3:Storing tweets to HDFS

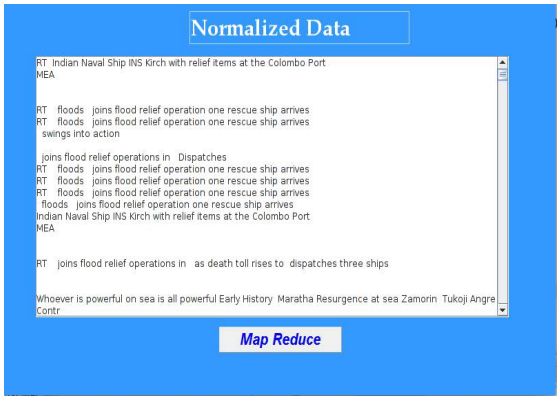

Figure 4: Normalized Tweets

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

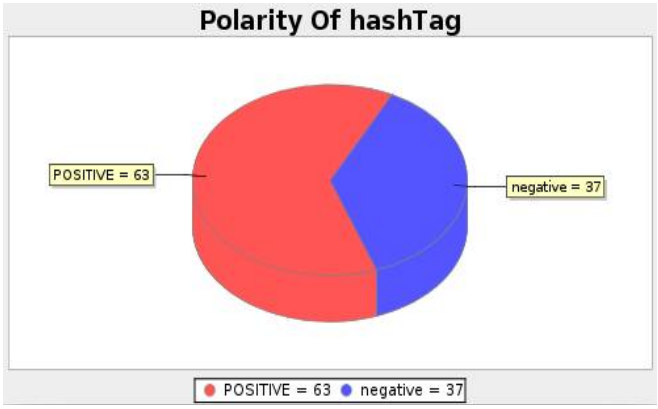

Figure 5: Output in the form of Pie Chart

VI. CONCLUSION AND FUTURE WORK

Big data is the term used for data sets that are hefty or complex that traditional data processing application software is inadequate to deal with them.A huge amount of data canbe stored and large computations can be done in a single compound with full safety and security at cheap cost.

The usefulness of the sentiment analysis for future marketing and businessby using a keyword and analysis of the sentiments around that keyword by the public will be more efficient if the speed of the processing can be increased.

REFERENCES

1. L.Jaba Sheela, “A Review of Sentiment Analysis in Twitter Data Using Hadoop”, International Journal of Database Theory and Application Vol.9, No.1 (2016), pp.77-86.
2. Ajinkya Ingle, Anjali Kante, Shriya Samak, Anita Kumari, “Sentiment Analysis of Twitter Data Using Hadoop”, International Journal of Engineering Research and General Science Volume 3, Issue 6, November-December, 2015.
3. Skuza, Michal, and Andrzej Romanowski. "Sentiment analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction" In Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on, pp. 1349-1354. IEEE, 2015.
4. Vasu Jain, “Prediction of Movie Success using Sentiment Analysis of Tweets”, The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3.
5. Lin, Jimmy, and Alek Kolcz. "Large-Scale Machine Learning at Twitter." In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp. 793-804. ACM, 2012.
6. Liu, Bingwei, Erik Blasch, Yu Chen, Dan Shen, and Genshe Chen. "Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier" In Big Data, 2013 IEEE International Conference on, pp. 99-104. IEEE, 2013.
7. AlvaroCuesta, David F., and María D. R-Moreno. "A Framework for Massive Twitter Data Extraction and Analysis", In Malaysian Journal of Computer Science, pp 50-67 (2014):1.
8. Bian, Jiang, Umit Topaloglu, and Fan Yu. "Towards Large-Scale Twitter Mining for Drug-Related Adverse Events" In Proceedings of the 2012 international workshop on Smart health and wellbeing, pp. 25-32. ACM, 2012.
9. Tare, Mohit, Indrajit Gohokar, Jayant Sable, Devendra Paratwar, and Rakhi Wajgi. "Multi-Class Tweet Categorization Using Map Reduce Paradigm" In International Journal of Computer Trends and Technology. pp 78 - 81 (2014).
10. T. Wilson, J. Wiebe and P. Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis,” in Proceedings of HLT and EMNLP. ACL, (2005), pp. 347–354.
11. R. Tushar and S. Srivastava, “Analyzing stock market movements using twitter sentiment analysis”, Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). IEEE Computer Society, (2012).
12. D. Pessemier and Martens “MovieTweetings: A Movie Rating Dataset Collected From Twitter”, Ghent University, Ghent, Belgium, (2013).