REAL-TIME OPINION MINING OF TWITTER DATA USING FLUME AND HADOOP

Sweta Mishra
M.tech Scholar
Department of Computer Science & Engineering
Bansal Institute of Science and Technology
Bhopal, India

Damodar Tiwari
Assistant Professor
Department of Computer Science & Engineering
Bansal Institute of Science & Technology
Bhopal, India

Shailendra Singh
Professor
Department of Computer Engineering Applications
National Institute of Technical Teachers Training and Research
Bhopal, India

Sanjeev Sharma
Professor & Head
Department of SOIT
Rajiv Gandhi Proudyogiki Vishwavidyalaya
Bhopal, India

***Abstract*— Today every internet user have a social account on which they share their opinions and communicate to each other. In all social sites Twitter is very popular on sharing the various user opinions towards a particular field. These twitter data is unstructured in nature so its hard to process with traditional tools. And also social websites like twitter , facebook, etc generates petabytes of data every day so storing these amount of data is also important. In these paper we proposed hadoop which is an open source framework used to stored and process huge amount of structured and unstructured data. In these we also propose various hadoop ecosystems which is used to fetch real time tweets and analyzing these data in efficient manner. Here we have used dictionary based approach for analysis for which we have implemented hive queries through which we can analysis these complex twitter data to check polarity of the tweets based on the polarity dictionary through which we can say that which tweets have negative opinion or positive opinion. Also we can perform some optimization to enhance query performance in hive.**

***Keywords*-- Hadoop, twitter, data mining, social analysis, hadoop ecosystem, flume, hive.**

I.INTRODUCTION

Micro blogging is a really well-liked communication tool used among net users. Twitter [15] is one amongst necessary social media website that receives scores of tweets a day on kind of important and trending problems. Users post their tweets on their life, share opinions on kind of topics and discuss current problems. These posts square measure then analysed by Government, Elections, Business, Product review etc. for higher cognitive process. Sentiment analysis is so, one amongst the vital space of study of twitter posts. Social media has gained Brobdingnagian quality among promoting groups, and Twitter is a good tool for an organization to urge individuals excited regarding its new merchandise launched. Twitter makes it simple to have interaction users and communicate directly with them, and successively, users will offer spoken promoting for corporations by discussing the products[11]. Given restricted resources, and knowing we tend to might not be able to consult with everybody we would like to focus on directly, promoting departments will be a lot of economical by being selective regarding whom we tend to reach intent on instead of effecting field surveys for getting feedback.

Sentiment Analysis on Twitter is harder than doing it for giant reviews. this can be as a result of the tweets square measure terribly short (only regarding a hundred and forty characters) and frequently contain slangs, emoticons, hash tags and alternative twitter specific jargon. For the event purpose twitter provides streaming API [17][20] that permits the developer associate degree access to a quarter of tweets tweeted at that point bases on the actual keyword. the item regarding that we would like to perform sentiment analysis is submitted to the twitter API's that will more mining and provides the tweets associated with solely that object. Twitter knowledge is usually unstructured i.e use of abbreviations is extremely high. Conjointly it permits the employment of emoticons that square measure direct indicator of the author's read on the topic. Tweet messages conjointly include a timestamp and also the user name. This timestamp is helpful for guesswork the long run trend application of our project. User location if

obtainable may facilitate to determine the trends in several countries.[6]

HADOOP

Hadoop [12] developers can deploy programs written in any other languages or in java for the processing of data parallely across multiple commodity machines despite of the fact that hadoop framework is written in java.
One of the key features of hadoop is that it partitions the computation and data across multiple nodes and then makes the application computation run in parallel on these nodes. Important features of hadoop are redundancy and reliability which means that if any of nodes fails due to technical fault or other failures, it automatically creates a backup for that node without any intervention of the operator.
Depending on the process complexity the time of execution may vary from minutes to hours. Hadoop has emerged out to be a potential solution for number of applications in web log analysis, visitor behaviour, search indexes, indexing and analysis of text content, applications in biology, genomics and physics, machine learning researches and natural language processing researches and in all sort of data mining.

Gathering Data with Apache Flume

To automate the movement of tweets from the API to HDFS, without our manual intervention, Flume is used. Apache Flume is a reliable and distributed system for effectively gathering and moving large amounts of data from various sources to a common storage area. Major components of flume are source, memory channel and the sink.

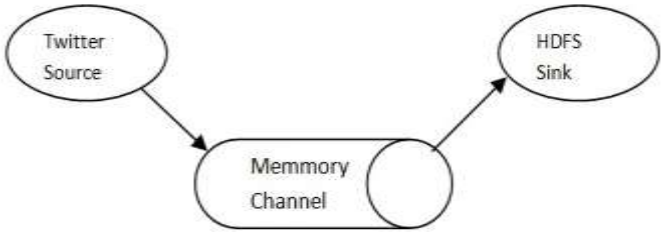

Figure 1. Architecture of Apache Flume

Twitter supply is associate degree event-driven supply that uses Twitter4j library for accessing streaming API. Tweets are collected and collective into elementary units of knowledge referred to as an occurrence. an occurrence incorporates a computer memory unit payload associate degreed an elective header. The coordination of event-flows from the streaming API to HDFS[14] is undertaken by Agent. The non heritable tweets are hold on into one or additional memory channels. A memory channel may be a temporary storage that uses associate degree in-memory queue to retain events till they're eaten by the sink. victimization memory channel, tweets are processed in batches which will be organized to carry a relentless range of tweets. to acquire tweets for a given keyword filter question is employed. Sink writes events to a pre organized location. this technique makes use of the HDFS-sink that deposits tweets into HDFS.

Hive

After assemblage the tweets into HDFS they're analyzed by queries victimization Hive. Apache Hive knowledge warehouse computer code facilitates querying and managing giant datasets residing in distributed storage. Hive provides a mechanism to project structure onto this knowledge and question the information employing a SQL-like language referred to as HiveQL. In opinion mining system, hive is employed to question out interested a part of the tweets which might be associate degree opinion, comments associated with a particular topic or a trending hash tag. Twitter API hundreds the HDFS with tweets that are portrayed as JSON blobs. process twitter knowledge in relational knowledgebase like SQL needs vital transformations as a result of nested data structures. Hive facilitates associate degree interface that has easy accessibility to tweets victimization HiveQL that supports nested knowledge structures. Hive compiler converts the HiveQL queries into map scale back jobs. Partition feature in hive permits tweet tables to separate into completely different directories. By constructing queries that has partitions, hive will confirm the partition comprising the result. the placement of twitter tables are expressly laid out in “Hive External Table” that are partitioned off. Hive uses SerDe (Serializer-Deserializer) interface in deciding record process steps. Deserializer interface intakes string of tweets and interprets it into a Java Object that Hive will manipulate on. The Serializer interface intakes a java Object that Hive has worked on and converts it into needed knowledge to be written on HDFS.

II. LITERATURE REVIEW

In [1] there's continuously a price to be learned from the period knowledge instead of historical knowledge from past twenty years. quite that we are able to even observe

a plague that is occurring at the instant. The results gift within the paper could be a complete open supply Piece of stories (PoN) answer for process and analyzing real-time streaming knowledge.[2]

In [3], massive knowledge have emerged in concert of the fascinating areas of analysis within the previous few years. The demand for analysis shows that it will grow within the next years to come back. massive knowledge primarily came into existence owing to the ascent of social media[5]. Twitter has looked as if it would be the one in every of the foremost well-liked social media over the web. Twitter receives tens of scores of tweets per day, making immense knowledge in unstructured kind, plenty of analysis has been administered to extract helpful info from twitter data. It additionally exhibits sentiment of the individuals on specific topics. However, this immense knowledge repository is unstructured and offers itself for several analysis areas. variety of researchers have tried to extract helpful info from this unstructured knowledge for varied applications. This paper presents a framework to check raw tweets in an exceedingly ascendable and optimum fashion. the most objective of the analysis work is to induce sentiment of the individuals and visualize it for higher understanding. Spring XD has been accustomed fetch tweets on a true time basis. These raw tweets area unit then reworked to Hadoop Distributed filing system (HDFS)[2]. Hadoop Scripting Language (HIVE) is employed to refine and label the tweets for his or her several sentiments. Finally, these sentiments area unit classified as positive, negative associate degreed neutral victimisation an formula that is simulated over HIVE. The proposed formula yields higher leads to term of sentiment.

In [7], they exploitation machine learning formula , a feature vector is constructed with the sensation describing words from tweets and are fed to the classifier that classifies the sentiment or opinion [16]. It aforesaid that varied twitter info analysis techniques that are supported lexicon that are exploitation the machine learning approaches.

In [8], large info [13] Sentiment Analysis exploitation Hadoop. the foremost focus of the analysis was to go looking out such how which can efficiently perform Sentiment Analysis on large info sets[9].Throughout in this Sentiment Analysis was performed on associate outsized info set of tweets exploitation Hadoop and so the performance of the technique was measured in form of speed and accuracy.

In [10], Associate in open supply methodology for predicting sentiments might alter North American country., within which they fetch the varied opinions from the online and predict customer's preferences. Till now, there are completely different problems predominating throughout this analysis community, namely, sentiment classification, feature based classification. During this they presents a review on techniques .

S.no.	Reference	Technique	Description
1	Nehal Mamgain, Ekta Mehta, Ankush Mittal, Gaurav Bhatt	Machine Learning, Neural Network	This paper is a thorough effort to dive into the novel domain of performing sentiment analysis of people's opinions regarding top colleges in India.
2	Kumar Chawda, Dr. Ghanshyam Thakur	Advanced Analytic Tool	This paper deals with the study of big data 5V's definition , Analysis requirement, tools, frame works and different type of cloud based big data analytics tools provide by different companies and functioning of Hadoop or MapReduce Process
3	Skuza, Michal, and Andrzej Romanowski	Artificial Neural Network	In this paper sentimental analysis is performed on the data extracted from Twitter and Stock Twits. The data is analyzed to compute the mood of user's comment.
4	Judith Sherin Tilsha S , Shobha M S	Ensemble Classifiers, Feature Vector	using machine learning algorithms were proposed in which a feature vector is constructed with the emotion describing words from tweets and are fed to the classifier that classifies the sentiment or opinion.
5	Ramesh R, Divya G, Divya D, Merin K Kurian	Machine learning, NLP	The main focus of the research was to find such a technique that can efficiently perform Sentiment Analysis on Big Data sets
6	G.Vinodhini , RM.Chandrasekaran	machine learning	This paper presents a survey covering the techniques and methods in sentiment analysis and challenges appear in the field.
7	Proposed work	Bigdata analytical tools	In this we can find perform the sentiment analysis on real time social data using open source solutions.

Table-1 Related work

Comparison with previous work

In [1], they perform real time sentiment analysis using machine learning algorithm called naive bayes for which they used python , which is very slow in terms of data fetching and i will take lots of time in data fetching and processing these data and as we know any social sites generates petabytes of data every data so that huge amount of data is fetched and stored by python is not possible. So we proposed an hadoop ecosystems which is very fast and accurate in terms of data gathering and storing large data into HDFS.

IV PROBLEM DEFINITION

Sentiment Analysis [4] is that the method of 'computationally' determinant whether or not a bit of writing is positive, negative or neutral. It's additionally

called opinion mining, derivation the opinion or angle of a speaker. it's necessary for

- Business: In promoting field firms use it to develop their methods, to grasp customers' feelings towards product or whole, however folks reply to their campaigns or product launches and why customers don't obtain some products.
- Politics: In political field, it's accustomed keep track of political read, to find consistency and inconsistency between statements and actions at the govt level. It is accustomed predict election results as well!
- Public Actions: Sentiment analysis is also accustomed monitor and analyse social phenomena, for the recognizing of probably dangerous things and determinant the final mood of the blogosphere.

Sentiment Classification Techniques

Their are two type of techniques is present for sentiment classification: lexican based(machine learning) and dictionary based approach. In lexicon based we can train the model and after training we can predict the sentiment based on the model learning capabilities. Another approach id dictionary based approach in which we can provide an external lexicon dictionary for finding sentiments. In this paper we also choosen dictionary based approach because it's has more accurate as compared to lexicon based approach.

Text mining help an organization derive potentially valuable business insights from text-based content such as word documents, email and postings on social media streams like Facebook, Twitter and LinkedIn. Data mining or Text mining plays a important role in decision making because through these mining techniques we can analyse the data and on the basis of result we can take a decision.

Sentiment classification is a technique to focus on the sentiments or opinions expressed in an article or conveyed orally. The term sentiment includes emotions, conclusions, behavior and others. In this paper, the work concentrates on human readable text writing on the e-commerce sites.

V. PROPOSED WORK

For analysing these large and complex data required a power tool, we are using hadoop[10] which is a open source implementation of mapreduce which is a powerful tool designed for deep analysis and transformation of very large data.

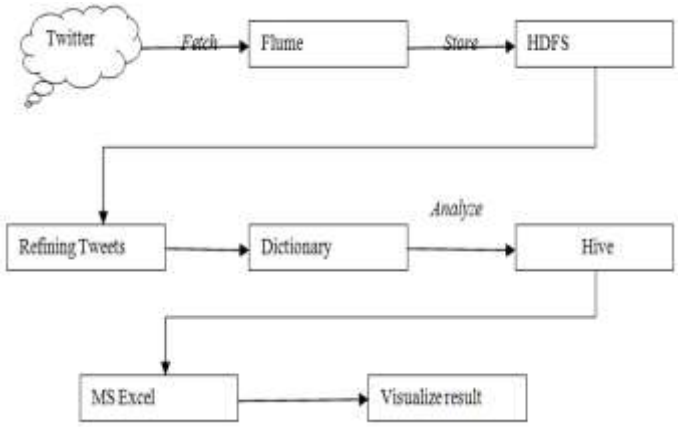

Figure-2. Flow Diagram of proposed work

Algorithm 1 Get Opinion as Positive, Negative and Neutral

Require: R - a raw tweets fetch from Flume
Require: D - a comprehensive dictionary to match the words in tweets
Ensure: TB - refine tweets with sentiment score

for all i to D do
if R[i]:Negative then
Score = -1
if R[i]:Positive then
Score = 1
else
score = 0
end if
end if
end for
for all i to D do
if R[i]:Polarity > 0 then
Sentiment = 1
if R[i]:Polarity < 0 then
Sentiment = -1
else
Sentiment = 0
end if
end if
end for
for all i to D do
if R[i]:Sentiment > 0 then
TB = 2
if R[i]:Sentiment < 0 then

TB = 1
else
TB = 0
end if
end if
end for

VI. EXPERIMENTAL & RESULT ANALYSIS

All the experiments were performed using an i5-2410M CPU @ 2.30 GHz processor and 4 GB of RAM running ubuntu 14 [10]. As we have seen the procedure how to overcome the problem that we are facing in the existing problem that is shown clearly in the proposed system. So, to achieve this we are going to follow the following methods:

- Creating Twitter Application.
- Getting data using Flume.
- Querying using Hive Query Language (HQL)

Creating Twitter Application

First we can create a twitter application through which we can get the access token keys, and consumer keys for authentication. Figure 3 shows the various access tokens keys which is used for making API [18][19] call from twitter to fetch the real time twitter data and stored it into the HDFS.

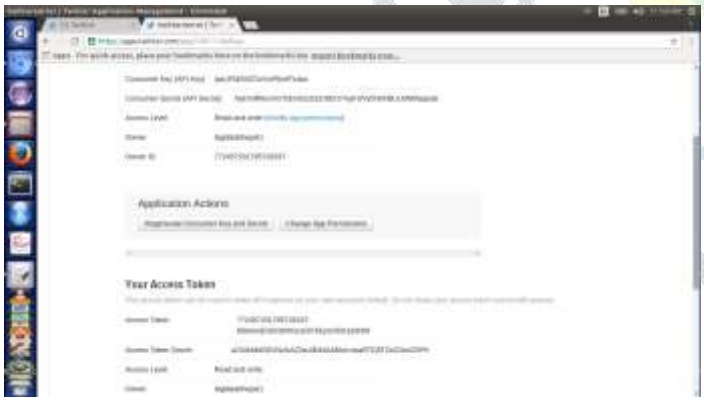

Figure-3. Creating twitter application & Generation access token keys

Getting data using Flume

After getting all the keys we can configure flume on top of the hadoop, and we can configure the following flume properties are shown below.

```
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS
Â
TwitterAgent.sources.Twitter.type =
com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = FlRx3d0n8duIQ0UvGeGtTA
TwitterAgent.sources.Twitter.consumerSecret =
DS7TTbxhmQ7oCUlDntpQQRqQllFFOiyNoOMEDD0lA
TwitterAgent.sources.Twitter.accessToken = 1643982224-
xTfNpLrARoWKxRh9KtFqc7aoB8KAAHkCcfC5vDk
TwitterAgent.sources.Twitter.accessTokenSecret =
PqkbuBqF3AVskgx1OKgXKOZzV7EMWRmRG0p8hvLQYKs
Â
TwitterAgent.sources.Twitter.keywords = hadoop, big data,
analytics, bigdata, cloudera, data science, data scientiest,
business intelligence, mapreduce, data warehouse, data
warehousing, mahout, hbase, nosql, newsql, businessintelligence,
cloudcomputing
Â
TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path =
hdfs://localhost:9000/user/flume/tweets/
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000
â
```

Querying using Hive Query Language

After fetching and storing the twitter data into HDFS, we can start analyzing these data by using apache Hive, for which we can create a table by writing hive query are shown below, and in the delimiters we can uses cloudera Json serde properties to validate the twitter unstructured data and transform these data into Structured form.

```
create external table raw(
created_at string,
id bigint,
text string,
user STRUCT<
screen_name:string,
name:string,
locations:string,
description:string,
created_at:string,
followers_count:int,
url:string>)
ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe'
location '/home/abhi/work/warehouse/new_bigdata';
```

After that we can load the data into the raw table and these serde properties validate the data and stored these data in to table raw. The data stored into raw table are shown in figure 4.

Figure 4. Data store into table raw

After that we can start refining these data , for which we can filter the whole data and fetch only the tweet id and tweets text for sentiment analysis , which is shown in figure 5.

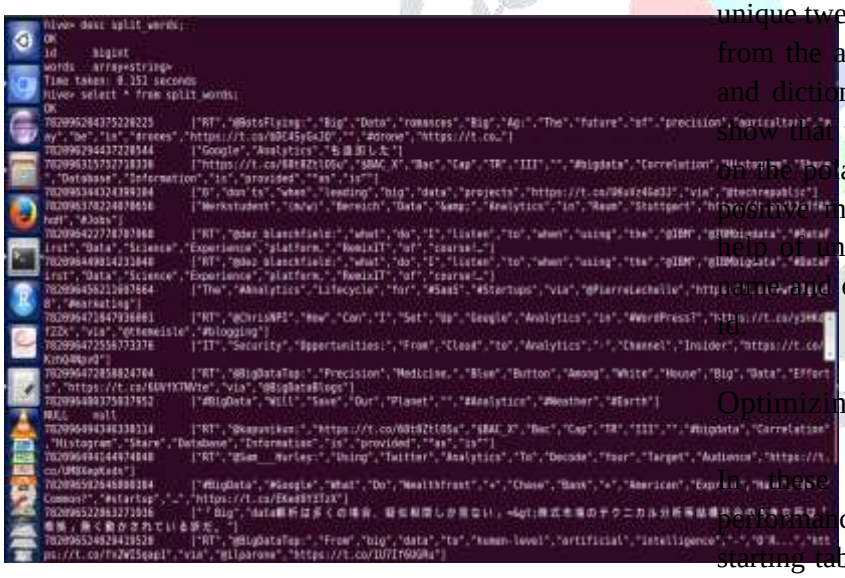

Figure 5. Data store into table word table

After that we can create a another table called dictionary table to store a dictionary which consist a two information first stored a English word and another field store a polarity of that tweet which is from -5 to +5. After that we have a two table , first table contains a two field which store a tweet id and the word related to that tweet and second table consist a two fields a English word and a polarity of that English word. Than we can join this two table based on the words which is common from both the table , we can perform a left outer join and the resultant table are shown in figure 6.

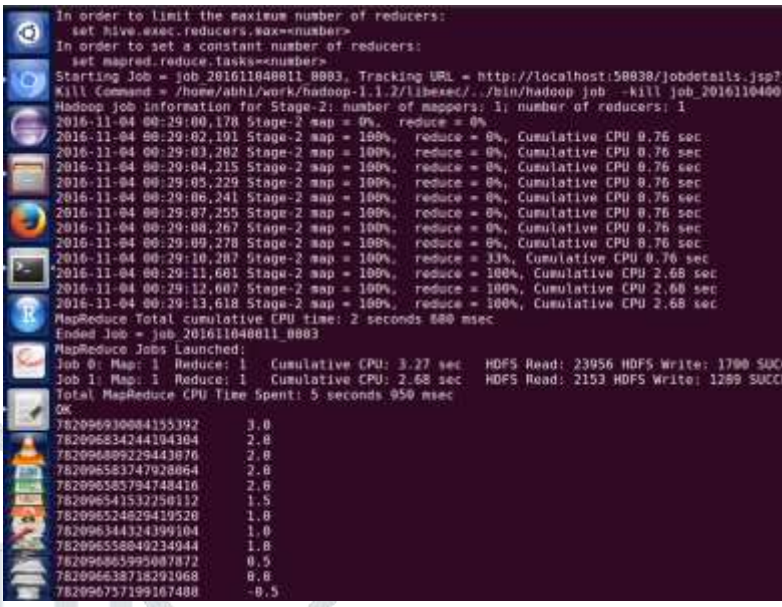

Figure 6. Data store in resultant table

The resultant table consist of a two fields which store a unique tweet id and a polarity of that tweets which comes from the average rating from the joining of split words and dictionary tables. The resultant table shown in fig. show that unique tweet id and the polarity of that tweet on the polarity basis we can say that the tweet indicates positive meaning or a negative meaning, and with the help of unique tweet id we can easily identify the user name and other information on the basis of unique tweet id.

Optimizing Query Performance

In these we can also optimize the hive query performance, we can perform serialization process at the starting table and store the resultant table into new table and then apply all the query on these new resultant table by which we can get the result faster as compared to perform same operation on deserialize table. For these we can execute the different query on two hive tables first table in which the optimization is not present and second in which we perform optimization process for which we can get output of queries with different execution time and the time taken by query in both the tables. For these we can create another table call lognew and the schema difference between both table are shown in figure 7.

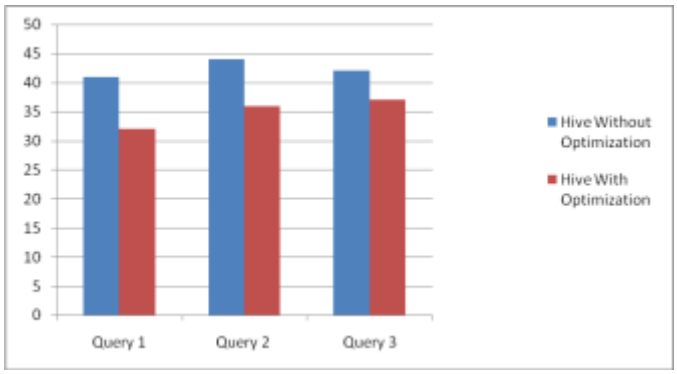

Figure 7. Execution time taken by query on hive table.

V CONCLUSION

Opinion Mining may be a terribly wide branch for analysis. we have lined a number of its necessary aspects. an equivalent design may well be used for a spread of applications designed to seem at Twitter knowledge, like distinguishing spam accounts, or distinguishing clusters of keywords. In this we can also identify the polarity of the tweet by which we can say that which tweet have a positive meaning or a negative meaning. In this we can also enhance query performance by optimizing hive query performance.

REFERENCES

[01] Ankur Goel, Jyoti Gautam, Sitesh Kumar, "Real Time Sentiment Analysis of Tweets Using Naive Bayes", in 2016 2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India 14-16 October 2016, IEEE.

[02] Can Uzunkayaa, Tolga Ensaria, Yusuf Kavurucu, "Hadoop Ecosystem and Its Analysis on Tweets" in World Conference on Technology, Innovation and Entrepreneurship, Elsevier 2015.

[03] Rashid Kamal, 2Munam Ali Shah, 3Asad Hanif, 4J Ahmad, "Real-time Opinion Mining of Twitter Data using Spring XD and Hadoop" in Proceedings of the 23rd International Conference on Automation & Computing, University of Huddersfield, Huddersfield, UK, 7-8 September 2017, IEEE.

[04] Manoj Kumar Danthala, "Tweet Analysis: Twitter Data processing Using Apache Hadoop", International Journal Of Core Engineering & Management (IJCEM) Volume 1, Issue 11, February 2015, pp 94-102.

[05] White Paper Big Data Analytics Extract, Transform, and Load Big Data with Apache Hadoop-Intel corporation.

[06] Judith Sherin Tilsha S , Shobha M S, "A Survey on Twitter Data Analysis Techniques to Extract Public Opinion", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 11, November 2015, pp 536-540.

[07] Ramesh R, Divya G, Divya D, Merin K Kurian , "Big Data Sentiment Analysis using Hadoop ", (IJIRST)International Journal for Innovative Research in Science & Technology,Volume 1 , Issue 11 , April 2015 ISSN : 2349-6010

[08] "Application Programming Interface." *Wikipedia*. Wikimedia Foundation, 23 Oct. 2014. Web. 24 Oct. 2014.

[09] G.Vinodhini , RM.Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey" , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012 ISSN: 2277 128X.

[10] Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More---Matthew A. Russell.

[11] K. W. Lim and W. Buntine, "Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon," in Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, ser. CIKM '14. New York, NY, USA: ACM, 2014, pp. 1319–1328.

[12] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," in the 26th IEEE Symposium on Mass Storage Systems and Technologies, pp. 1-10, May 2010.

[13] Sagiroglu, S., & Sinanc, D, "Big data: A review", IEEE International Conference on Collaboration Technologies and Systems (CTS), 2013, pp 42-47.

[14] Mirko Lai, Cristina Bosco and Viviana Patti, Daniela Virone, “Debate on Political Reforms in Twitter: A Hashtag-driven Analysis of Political Polarization” in IEEE, 978-1-4673-8273-1/15, IEEE 2015.

[15] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a Social Network or a News Media?", In: Proceedings of the 19th International Conference on World Wide Web, 2010, pp. 591-600.

[16] M. Cha, H. Haddadi, F. Benevenuto, and K.P. Gummadi, “Measuring User Influence in Twitter: The Million Follower Fallacy”, In: Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, 2010.

[17] "Application Programming Interface." *Wikipedia*. Wikimedia Foundation, 23 Oct. 2014. Web. 24 Oct. 2014.

[18] "Twitter's API --- HowStuffWorks." *HowStuffWorks*. N.p., n.d. Web. 24 Oct. 2014.

[19] Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More---Matthew A. Russell.

[20] "The Streaming APIs." *Twitter Developers*. N.p., n.d. Web. 23 Oct. 2014.