Analysis of Twitter Data Using Hadoop Ecosystem

R.Sindhuja[1], K.Manimekalai[2]
[1]Assistant Professor, Dept of Computer Applications
[2]Assistant Professor & Head, Dept of Computer Applications
[1, 2] Sri GVG Visalakshi College For Women, Udumalpet.

Abstract- *Twitter is one of the renowned social media that gets a huge amount of tweets every day. This information can be used for economic, industrial, and social or government approaches by arranging and analyzing the tweets as per our demand. Hadoop is one of the framework in Big data which is used to process different varieties of data using Hadoop Ecosystem. This paper discusses how the hadoop ecosystem mainly Apache Hive, Pig and Flume will be useful for Twitter data analysis.*

Keywords- FLUME, Hadoop, Hive, HDFS, Pig, Twitter, Yarn.

I. INTRODUCTION

Now a day, the data on the internet is growing in the rapid pace. Social media plays a vital role for the generation of data. These are numbers generated every minute of the day:[13]

- Snapchat users share 527,760 photos
- Users watch 4,146,600 YouTube videos
- 456,000 tweets are sent on Twitter
- Instagram users post 46,740 photos

Twitter is one of the micro bloggingsites which receives millions of tweets every data. The tweets contain the history of the author, opinions on any topics, discussions about current issues etc. Those tweets are used for decision making in various areas like Elections, government, Marketing. Sentiment Analysis is area which is used to analyze the post of twitter data which will be useful for decision making.

Analysis on twitter data is little complicated because the tweets are very short, which is of 140 characters. That tweet contains emoticons, hash tags, slang and some other specific jargon. The data which are generated can be in the form of Structured or unstructured. The data which is generated in the Microblogging platform can be analyzed using Hadoop Ecosystem.

II. LITERATURE REVIEW

In the past years, many works have been published in Sentiment Data Analysis. They have used different techniques to analyse the data from Twitter. There exist many possible variants. Some of them are discussed in this section.

In the year 2017, ShubhamGoyal implemented KNN Algorithm and Naive Bayes Algorithm to analyse the sentiment data of Twitter by extracting the data using Twitter API.[8]

Tao Chen et al proposed a divide and conquer approach which first divide the sentences into different types and then performs the sentiment analysis separately. To extract the target expression they used an approach called BiLSTM-CRF. And finally the sentiments of the data are classified using 1d-CNNs. This technique boosts the performance of sentence-level sentiment analysis.[12]

In the year 2016, Sangeetha who proposed two Ecosystem namely FLUME and HIVE from Hadoop Ecosystem to analyze the Twitter Data. For doing twitter analysis first data is collected using FLUME in local HDFS then the data is processed using the data warehouse infrastructure tool called Hive. [2]

Nikolaos Nodarakis et al analysed the Large Scale Sentiment data using Spark in the year 2016. They developed a novel distributed algorithm implemented in Spark which translates the programs into MapReduce jobs. They utilized Bloom filters to compact the storage size of intermediate data and boost the performance of the algorithm.[3]

Ramesh R et al analysed the data by splitting it into three levels, such as Document level, Sentence level and Aspect/Entity level in the year 2015. They used the algorithm named Sentiment Calculation Algorithm which was done on every tweet and a polarity score is given to it. The polarity score is calculated by using mapreduce programming model. This approach achieved the accuracy of 75%.[4]

PeimanBarnaghi et al used one of the well-known machine learning methods for text categorization and determining sentiment, called Logistic Regression Classification (LRC). Using that method they analysed the text and Sentiment polarity on FIFA World Cup 2014 Tweets.[11]

In the year 2014, Sunil B.Mane et al proposed the system of splitting the modules of data into few steps like Real time data and features, Part of speech, Root form, Sentiment Directory, Map reduce algorithm. In the final step they used a standard algorithm called PMI-IR 2 and achieved an accuracy of 72.27%.[1]

PrabhuPalaniswamy et al who used a simple lexicon based technique to extract sentiments from twitter data in the year 2013. They used the technique named Serendio taxonomy for analysing the sentiment of data and achieved the F-Score of 0.8004 on the test data set.[5]

Liu et al implemented Navie Bayes Classifier to achieve fine-grain control of the analysis procedure for a Hadoop implemention which resulted in a 80.85% average accuracy in the year 2013.[13]

III. HADOOP

Apache Hadoop is a best choice for analysing twitter data because it works for distributed big data. Apache Hadoop is one of the open source framework which is written in java.Hadoop consists of three key parts –

- **Hadoop Distributed File System (HDFS) –** It is the layer where storage of data will be performed.
- **Map-Reduce** – It is the layer where processing of data will be accomplished.
- **YARN** – It is the layer where the management of Hadoop will takes place.

Hadoop works in **master-slave** fashion. There is one master node and there can be n number of slave nodes, where n can be of any number. Master node monitors, process and manages the slave nodes. Slave node will perform the actual work.

Hadoop Framework includes different modules like MapReduce, YARN, Hive, Pig, HBase, HCatalog, Thrift, Apache Mahout, Apache Sqoop, Apache Flume, Zookeeper, Oozie for different functionality is shown in the below diagram.

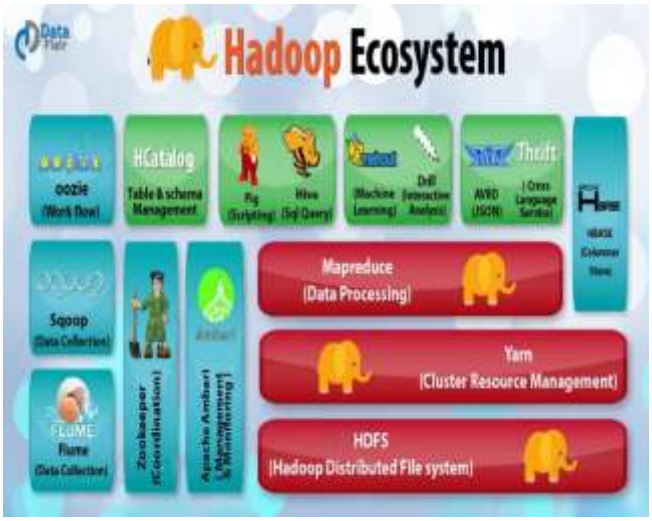


Fig 1: Apache Hadoop Ecosystem

Hadoop MapReduce is a framework used for processing large amount of data. Data are retrieved from varies source and stores in HDFS, which provides the way to perform parallel processing.

IV. HADOOP ECOSYSTEM

There are several tools used for sentiment data analysis. They are:

Hadoop Distributed File System(HDFS): It is a primary storage system of Hadoop. The data which is collected from the source like twitter will be stored in HDFS for processing. It contains NameNode(Master node) which stores only meta data and DataNode(Slave node) which stores the actual data.

Map Reduce: It is a software framework for easily writing applications that process the vast amount of structured and unstructured data stored in the Hadoop Distributed File system.
There are two Phases in this Ecosystem, One is Map Phase and another is Reduce Phase. Each phase has **key-value pairs** as input and output.[15]

YARN: It is called as the operating system of Hadoop as it is responsible for managing and monitoring workloads.It allows multiple data processing engines such as real-time streaming and batch processing to handle data stored on a single platform.[15]

Hive: It is an open source data warehouse system for querying and analyzing large datasets stored in Hadoop files.Hive do three main functions*:* data summarization, query, and analysis.*[15]*

Pig: Pig as a component of Hadoop Ecosystem uses ***PigLatin*** language. It is very similar to SQL. It loads the data, applies the required filters and dumps the data in the required format.

HBase: It is a Hadoop ecosystem component which is a distributed database that was designed to store structured data in tables that could have billions of row and millions of columns.

Apache Flume: It efficiently collects, aggregate and moves a large amount of data from its origin and sending it back to HDFS.

HCatalog: It is a key component of Hive that enables the user to store their data in any format and structure.

Avro: It is an open source project which provides data serialization and data exchange services for Hadoop. These services can be used together or independently.[15]

Thrift: It is a software framework for scalable cross-language services development. Thrift is an interface definition language for RPC(Remote procedure call) communication.[15]

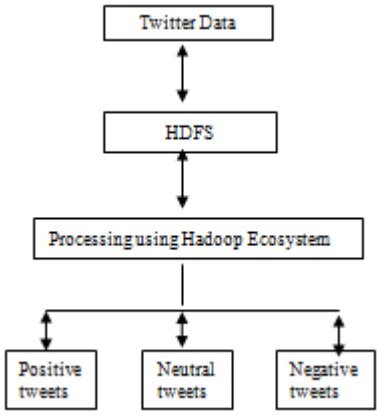
Apache Drill: The drill is the first distributed SQL query engine that has a schema-free model.

Apache Mahout: It is open source framework for creating scalable **machine learning** algorithm and data mining library. Once data is stored in Hadoop HDFS, mahout provides the data science tools to automatically find meaningful patterns in those big data sets.

V. PROPOSED SYSTEM

The main objective of this analysis is,

- Data Collection– The twitter dataset can be retrieved via twitter API.
- Data Storage – The tweets which are retrieved can be stored in Hadoop Distributed File System (HDFS).
- Processing of Data – Most of the data which are collected can be of unstructured; those are processed using Hadoop Ecosystems like Hive, Map Reduce.
- Data Analysis - The output obtained from reducer phase is analysed.
- Data Representation– Data can be represented using different methods such as pie chart.
- Output–The output of the above analysis is positive, negative and neutral tweets.



To process the analysis, Apache Flume which is used to collect, aggregate and moves large amount of data and send it back to HDFS. Other Component called Hive or Pig are used to analyse the twitter data.Analysis of twitter data consists of few steps:

Tokenization: All the words in tweets are broken into token to make the processing easy. The process of this conversion is called Tokenization.

For example: "@Karthick, The movie is veryyyyyy NICE" is broken down into tokens such as '@Karthick', 'the', 'movie', 'is', 'veryyyyy', 'NICE'. Emoticons, abbreviations, hashtags and URLs are recognized as individual tokens.

Normalization: The process of verifying and computing the tokens based on the kind of token is called as Normalization.

- If the token is an acronym, it is checked in the acronym dictionary and the corresponding full form will be replaced.
- If the token is in upper case, then it will be converted into lower case and stored.
- If the token is an emoticon, then its corresponding polarity is checked using emoticon dictionary.
- Spelling of character repetitions such as 'veryyyyy' are first corrected into 'very' and then stored as 'very'.
- The normalization process also discards all those tokens which, in no way, contribute to the sentiment of a tweet such tokens are called stop word. The URL will also be discarded.

The above analysis will be helpful for determining the reviews in the twitter as positive, negative or neutral.

VI. CONCLUSION

Sentiment Data Analysis is a wide area for research. Twitter is widely used social media for sharing the opinion on different issues and topics. Analysis of twitter data will be helpful for making the decision. Apache Hadoop is one of the tool used for twitter data analysis. Hadoop Ecosystem such as FLUME, pig and Hive are mostly used for analysing the twitter data. The outcome of the analysis is finding the polarity of tweets collected (Positive, Negative, Neutral). It can be useful for finding the reviews of a film, finding the mood of people while election.

REFERENCES

[1] Sunil B. Mane , Sunil B. Mane, YashwantSawant, SaifKazi, VaibhavShinde , "Real Time Sentiment Analysis of Twitter Data Using Hadoop", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3098 – 3100 , ISSN:0975-9646.

[2] Sangeeta,"Twitter Data Analysis FLUME & HIVE on Hadoop Framework" (IJRAET) International Journal of Recent Advances in Engineering & Technology, V-4 I-2 , 2016, ISSN (Online): 2347-2812.

[3] Nikolaos Nodarakis, Spyros Sioutas, AtanasiosTsakalidis, Giannis Tzimas, "Large Scale Sentiment Analysis on Twitter with Spark" published in the workshop Proceedings of the EDBT/ICDT 2016, ISSN: 1613-0073.

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

[5] PrabhuPalanisamy, Vineet Yadav, HArshaElchuri, "Serendio: Simple and Practical lexicon based approach to sentiment analysis", Serendio Software Pvt Ltd, 2013.

[6] Mahalakshmi R, Suseela S, "Big-SOSA: Social Sentiment Analysis and Data Visualization on Big Data", International Journal of Advanced Research in Computer and Communication Engineering, Vol.4, Issue 4, April 2015, ISSN:2278-1021.

[7] G.Vinothini, RM. Chandransekaran, "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.

[8] ShubhamGoyal, "Review Paper on Sentiment Analysis of Twitter Data Using Text Mining and Hybrid Classification Approach", IJEDR 2017, Vol 5, Issue 2, ISSN: 2321-9939.

[9] M. Vadivukarassi, N. Puviarasan and P. Aruna, "Sentimental Analysis of Tweets Using Naive Bayes Algorithm", World Applied Sciences Journal 35 (1): 54-59, 2017, ISSN 1818-4952.

[10] Ajinkya Ingle, Anjali Kante, ShriyaSamak, Anita Kumari, "Sentiment Analysis of Twitter Data Using Hadoop", International Journal of Engineering Research and General Science Volume 3, Issue 6, November-December, 2015, ISSN 2091-2730.

[11] PeimanBarnaghi, ParsaGhaffari, John G. Breslin, " Text Analysis and Sentiment Polarity on FIFA World Cup 2014 Tweets" Conference ACM SIGKDD'15, August 10-13, 2015, Sydney, Australia.

[12] Tao Chen ,RuifengXu,Yulan He , Xuan Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN".

[13] Bingwei Liu, Erik Blasch, Yu Chen, Dan Shen and GensheChen,"Scalable Sentiment Classification for Big data analysis using Naive Bayes Classifier",IEEE Intl Conf. On Big Data, Oct 2013,

[14] https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#6e9224ef60ba.

[15] https://data-flair.training/blogs/hadoop-ecosystem-components/

BIOGRAPHY OF THE AUTHORS

Sindhuja R has completed her MCA in Sri Ramakrishna Engineering College, Coimbatore and BCA degree in Sri GVG Visalakshi college for women, Udumalpet. She is a Rank Holder in BCA and MCA. At present she is working as an Assistant Professor of Computer Applications in Sri GVG Visalaskshi College for Women, Udumalpet, Tamilnadu, India. Her area of interest is Big Data Analytics.

Manimekalai K has completed her MCA and M.Phil degree. She has published 8 International journals. She has 14 years of Teaching experience and qualified State Level Eligibility Test(SLET). At present she is working as a Head & Assistant Professor of Computer Applications in Sri GVG Visalaskshi College for Women, Udumalpet, Tamilnadu, India. Her area of interest is Data Mining.