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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Sentiment Analysis on High Dimensional Data using Hadoop

Amit K. Burde[1], Vikram B.Shevate[2], Tushar Barage[3], Preeti Suryawanshi[4], Prof.Viresh Chapte[5].

UG Student, Dept. of Computer, DYPSOET, Lohegaon, India[1]

UG Student, Dept. of Computer, DYPSOET, Lohegaon, India[2]

UG Student, Dept. of Computer, DYPSOET, Lohegaon, India[3]

UG Student, Dept. of Computer, DYPSOET, Lohegaon, India[4]

Assistant Professor, Dept. of Computer, DYPSOET, Lohegaon, India[5]

ABSTRACT: Sentiment analysis is very popular now days. Identification of emotion of text form web source can be done using Sentiment Analysis system. Sentiment analysis also called opinion mining aims to find the polarity of text, which can be taken from any source. Most of these sources are social networking sites and Ecommerce sites. This project aims to use the advantages of Hadoop for Distributed storage and Distributed processing as well. Twitter Dataset will be retrieved using the Flume which will help string that vast in the HDFS. MapReduce algorithm will be used for the processing and the output or result of Sentiment analysis will be shown visually using the graph.

KEYWORDS: Hadoop, Sentiment Analysis, MapReduce, HDFS

I. INTRODUCTION

A. Sentiment Analysis

In simple words, process of Sentiment analysis is the task of classifying whether the sentiments, thoughts, reviews posted in a text is showing positivity or negativity about product, service or person. Sentiment Analyses focuses on dividing text by their opinion and emotions expressed by the person. Finding a text's polarity as positive or negative is a two-class problem. This is known as sentiment orientation analysis in text classification process. Sentiment analysis has been being very helpful. With the rapid development of the internet, web and smart devices users are always posting their opinions on the internet. feedback and comments based on those gives the marketing people improve their product and services. So Sentiment analysis is being used as a great marketing tool. Importance of sentiment analysis is increasing greatly.

B. Hadoop

Hadoop is an open-source framework that permits to store and process enormous information in a distributed domain crosswise over bunches of PCs utilizing basic programming models. It is intended to scale up from single servers to thousands of machines, each intended for distributed computing and distributed storage. Hadoop have two major component HDFS and MapReduce

a. MapReduce

MapReduce is a distributed computing technique based on java. The MapReduce algorithm is a set of two task Map and Reduce. In Map stage a set of data is taken which is then converted it into another set of data, where every single entity is broken into set of tuples (key/value pairs). Then the, reduce stage which takes the output from a map stage as an input and then the combining of those data tuples into a smaller set of tuples is done.After processing these tuples, it produces a new set of output, which will be stored in the HDFS.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

b. ***HDFS***

HDFS holds extensive measure of information and gives simpler access. To store such enormous information, the records are put away over various machines. These documents are put away in excess design to protect the framework from conceivable information misfortunes if there should be an occurrence of data loss. HDFS likewise makes applications accessible to parallel processing.

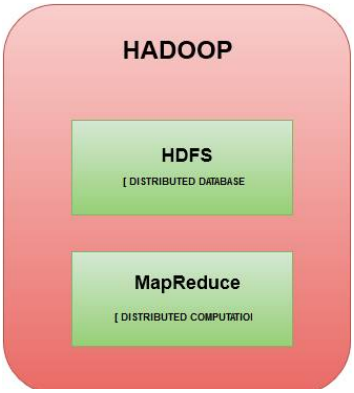

Figure 1 : Hadoop Architecture

C. Flume

Flume is a standard, basic, powerful, adaptable, and extensible tool for information ingestion from different web servers into Hadoop. Flume is utilized to move the log information created by application servers into HDFS at a higher velocity. Alongside the log records, Flume is additionally used to import enormous volumes of information created by Facebook and Twitter, also E-Shopping sites like Amazon.

D. Tokenization

Tokenization of text involves splitting of text into words, sentence or tokens

Example: Battery is Bad.

Above sentence after tokenization will be split**.** After tokenization we will get tokens “Battery”,”is”,”Bad” and ”.”

E. Stop Word Removal

These are the most common English words and their importance in search query or finding polarity of sentence is very minor. Their removal is necessary for cleaning out the unwanted text from processing.

***Example* :**because, about, a, an, the

F. Porter Stemming

Stemming is the process of bringing the word in to its original root from.

***Example* :** Stemming , Stemmed, => Stem

G. Negation Handling

These are the words which alters the meaning of sentence from positive to negative and vice versa. So handling of these negative words is very important in Sentiment Analysis

***Example* :**No, Never, Neither, Nor

H. TF/IDF

a. **TF(Term Frequency)** is identifying the frequency of particular terms in the sentence
TF(t) = (frequency of term t in a document) / (Total terms in the document).

b. **IDF(InverseDocumentFrequency)** is identifying the importance of the term in that sentence.
IDF(t) = log_e(total documents / no of documents with term t).

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

II. RELATED WORK

Sentiment analysis is exceptionally prevalent these days. Heaps of research is going around the natural language processing information extraction and data mining. Many researchers are exploiting different methods and tools to obtain the sentiment analysis. Any system must have good speed and efficiency too. Without the speed and efficiency, a system might not be suitable for performing the sentiment analysis on the vast amount of data that is used for the opinion mining.Sunil B. Mane and his colleagues provided [1] efficient way of doing sentiment analysis using Hadoop. Their approach was focused on the speed. Technique provided by them processed vast amount of data on a Hadoop cluster faster in real time.JayshreeKhairnar[2] discussed about using Support Vector Machin and LSA to perform sentiment analysis. They found that there is still need of improvements in terms of efficiency and accuracy. Hence the reason they suggested to use MapReduce. Their method gave better efficiency for producing the result.Jeffrey Shafer[3] discussed the causes of performance bottlenecks in Hadoop. He identified three different problems like software architecture bottlenecks, portability limitation and portability assumption. Jeffery found that the problem of bottlenecks is more related to HDFS than to MapReduce.Sentiment analysis result can be shown using different types of graphs and charts**.** Changbo Wang presented another[4] representation framework for doing analysis and visualizing and verifying the sentiments from web source. In SentiView, different methods of visualization to show the output of sentiment analysis have been added. Ruchika Sharma proposed a system[5] in which her emphasis was on improving the accuracy of sentiment analysis. Her method of using Multiple kernel gave better accuracy of 90% and 92%. Though her conclusion was of using multiple kernel with some different Machine leaning algorithms does not improve on accuracy. DhirajGurkhe[6] tried different datasets for his proposed methods. Datasets were mainly from the social networking sites. Dhiraj Concluded that his system gives best results with Unigram detection.

III. PROPOSED SYSTEM

A. *Proposed System Architecture*

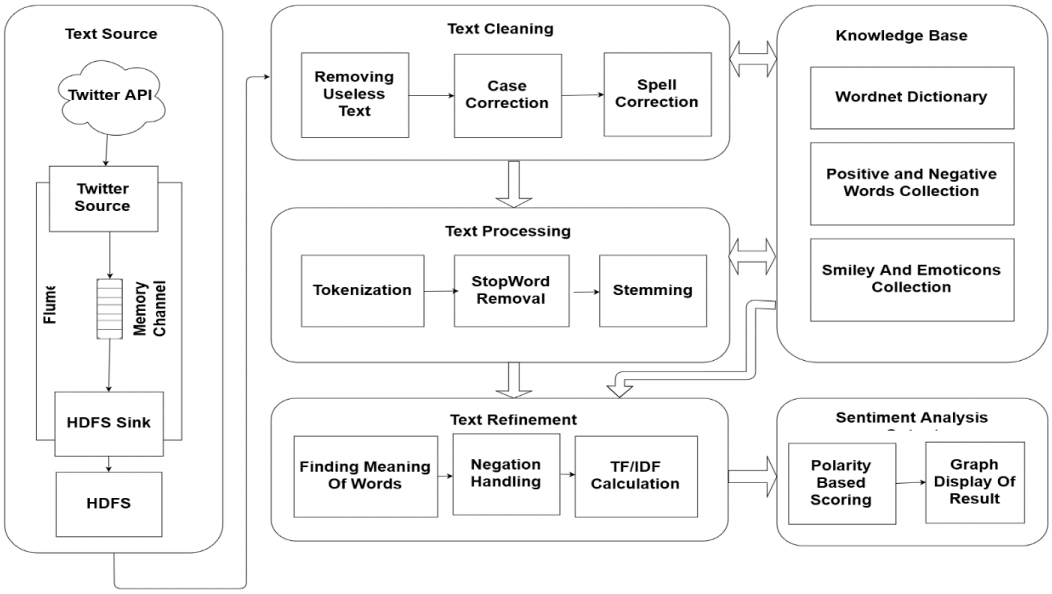

Figure2: System Architecture

B. *Description of Architecture*

Sentiment analysis starts with taking a text source on which the sentiment analysis is going to be performed. With the help of Twitter API and Flume database of twitters is transferred to the HDFS sink which is then stored to HDFS. Knowledge Base in this architecture contains the WordNet Dictionary. This WordNet Dictionary contains all the

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

English word out there with their meanings and true and negative ratio. Different smiley and emoticons are also added. Knowledge Base is used by both the text cleaning and text processing operations. Different cleaning operations are done on the tweets. '#' and Urls are removed. Case corrections and spell corrections are done to clean the text. Text processing is started with first tokenization being done. Stop Words are then removed as they don't have any meaning in terms of identifying the negativity or positivity. Stemming is done to remove the suffixes from words. Text refinement involves most important task of handling negations. Negations can make the result wrong so they must be handled carefully. TF/IDF calculation is used to identify term frequency and the importance of frequency. Finally, the Scores are calculated on the basis of polarity of Positive and Negative and the results are then displayed using the Graph.

IV. EXPERIMENTAL RESULTS

In this area, we show experiments and their results. We performed different task like acquiring data from twitter, pre-process data to remove noise from it, processing of that data then polaritydetection and finally producing graphs to show the result visually.Sentiment analysis was done on the tweets or comments from web source taken using flume. Using flume, the tweets were stored in the HDFS. From there the tweets wereprocessed by using MapReduce. Different text cleaning operations were done to get more useful data. Tokenization was done splitting of sentence in the tokens. These tokenized words were then stemmed to get their root form. WordNet Dictionary was used find the meaning of words and get the weight of those words in terms of positive and negative ratio. Finally, the score was calculated and the result was produced and the result was shown visually using Graph.

Pos	ID	Positive Score	Negative score	SynSetTerms
N	05144079	0	0.875	bad#1
N	04152593	0	0	Screen#1

Table 1: SentiWordNet Scoring System

While doing the experiment accessing time and processing time was very minimum due to use of better hardware and Haddop Framework. Table 1 shows the SentiWordNet scoring of positive or negative ratio. Most of the stopwords are not present in the SentiWordNet dictionary. Since stopwords are removed already only calculation of meaningful words/terms will be done. This also gives better time efficiency. Figure 3 Shows how many Negative or Positive words were there in the data set.

Parameter	Total	Positive	Neutral	Negative
Words	9457	3625	1657	4175
Percentage	100	38.33	17.52	44.14

Table 2 : Result

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

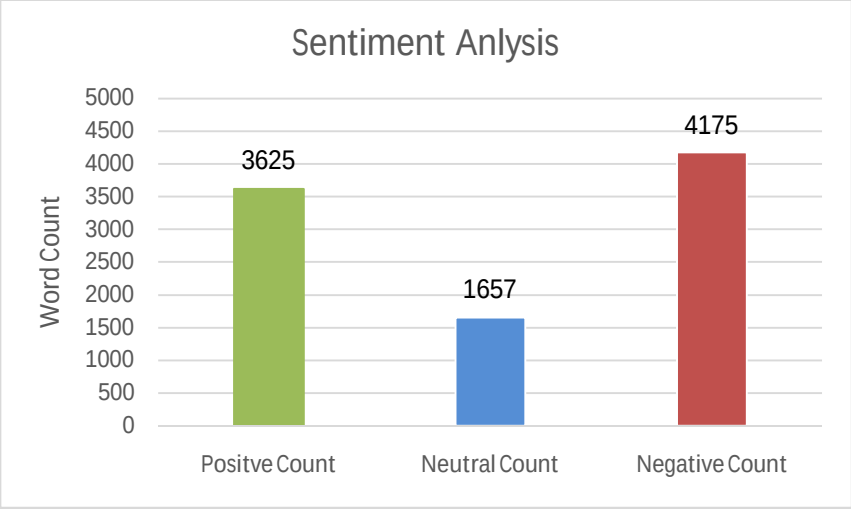

Figure 3 :Negative and Positive Words Count

V. CONCLUSION AND FUTURE WORK

There is lot of potential research to be done in the sentiment analysis. We tried to cover most of the important aspects of the sentiment analysis. Hadoop is well known for its advantages for distributed computing and distributed database. Flume makes it possible for getting that vast amount of tweeter data using twitter API. Getting the tweets closely related to the aspects can give better data sets.Negation handling is very important in sentiment analysis and should have better output. Since the negation handling can really impact the result of sentiment analysis.Future work can be done on implementing different language dictionaries as current approach supports only English Language. Finding different emotions like angry, Boring, Exciting from sentiment analysis can also be considered. Different ways to produce the output visually can be thought of too.

VI. ACKNOWLEDGEMENT

I would like to thank all the experts for providing their expertise for this research would like to thank my family for encouraging to me for doing always as best as I can. I really appreciated the guidance provided by Prof. BhagyashreeDhakulkar, Mr. Pravin Jadhav, Mr. SandipGharate, Mr. JavedTamboli. I loved working with my team, they were the best. I would like to say thanks to all my friends for any help they provided.

REFERENCES

1. Sunil B. Mane et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3098 – 3100
2. JayashriKhairnar, MayuraKinikar “Sentiment Analysis Based Mining and Summarizing Using MapReduce” 27th IEEE Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2015, 4081-4085
3. Shafer, Jeffrey, Scott Rixner, and Alan L. Cox. "The hadoop distributed filesystem: Balancing portability and performance." *Performance Analysis of Systems & Software (ISPASS), 2010 IEEE International Symposium on*. IEEE, 2010.
4. Wang, Changbo, et al. "SentiView: Sentiment analysis and visualization for internet popular topics." *Human-Machine Systems, IEEE Transactions on* 43.6 (2013): 620-630.
5. Ruchika Sharma and Amit Arora. Article: Improve Sentiment Analysis Accuracy using Multiple Kernel Approach. *International Journal of Computer Applications* 71(20):12-15, June 2013. Full text available
6. DhirajGurkhe, Niraj Pal and Rishit Bhatia. Article: Effective Sentiment Analysis of Social Media Datasets using Naive Bayesian Classification. *International Journal of Computer Applications* 99(13):1-4, August 2014.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

BIOGRAPHY

Amit K. Burde is undergraduate student in the Dr. D. Y. Patil School of Engineering and Technology, Lohegaon, India. He is pursuing Bachelors of Engineering in Computer Science. His Research area interests include Parallel Computing, Big Data, Computer Networks and Algorithms.

VikramShevate is undergraduate student in the Dr. D. Y. Patil School of Engineering and Technology, Lohegaon, India. He is pursuing Bachelors of Engineering in Computer Science. His Research area interests include Computer Networks and Computer Security.

TusharBarage is undergraduate student in the Dr. D. Y. Patil School of Engineering and Technology, Lohegaon, India. He is pursuing Bachelors of Engineering in Computer Science.His Research area interests include Computer Networks and Computer Security.

PreetiSuryawanshi is undergraduate student in the Dr. D. Y. Patil School of Engineering and Technology, Lohegaon, India. She is pursuing Bachelors of Engineering in Computer Science. Her Research area interests include Computer Networks.

VireshChapte is Assistant Professor in Dr. D. Y. Patil School of Engineering and Technology, Lohegaon, India.