International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 4
ISSN: 2454-4248
576 – 578

Sentiment Analysis of Twitter Data

[1]Mr. Satish Kumbhar, [2] Ashwini Lohar, [3]Pratiksha Sheth, [4] Sayli Talathi
[1]Assistant Professor, Department of Computer Engineering, College of Engineering, Pune, Maharashtra 411005
[2]B.Tech Student, Department of Computer Engineering, College of Engineering, Pune, Maharashtra 411005
[3]B.Tech Student, Department of Computer Engineering, College of Engineering, Pune, Maharashtra 411005
[4]B.Tech Student, Department of Computer Engineering, College of Engineering, Pune, Maharashtra 411005
Email: [1]ssk.comp@coep.ac.in, [4] saylitalathi1@gmail.com
Contact:[1] +919860574798,[2] +919503541418, [3]+918412072854, [4] +918390355296

Abstract— Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, emotions, political and religious views from written language about personality, product or event and determined whether they are viewed positively or negatively. Our project will involve collection of data from web resources such as twitter by using Hadoop and intend to derive useful inferences and recommendations. From the evaluation of this study it can be concluded that the proposed machine learning and natural language processing techniques are an effective and practical methods for sentiment analysis.

Keywords*-opinion Mining, Hadoop, Natural Language Processing.*

*****

I. INTRODUCTION

Over last 10 years 'BIG DATA' has growing importance in various industries, and generates huge data every day. The term Big Data is used for sets of huge size like the traditional databases that are unable for processing operations in less amount of time. As data increases the main challenge is storing huge amount of data, accessing and analyzing it in less amount of time. Hadoop is used to solve this problem. Hadoop is open-source implementation of the MapReduce programming for storing as well as processing big data and it is also scalable .We also know that many different industries and survey companies take decisions by collecting data from the web. WWW(World Wide Web) contains unstructured data. For analysis of that data we can collect it based on specific situation or on a distinguished thing.

Twitter is an one of the social media which has rich amount of data that can be a semi-structured, unstructured and structured data. Proposed system has effective sentiment analysis done on the data which is collected from the Twitter's API. Further, analysis is done using different machine learning algorithms such as naïve bayes, SVM, Decision Trees, etc.

II. TOOLS

Hadoop**:**

Apache Hadoop is a distributed framework which is good for sentiment analysis. It has different components like MapReduce, Flume, Hive, Pig, Sqoop, Oozie, Zookeeper, Hbase .HIVE and FLUME is used for sentiment analysis. Hadoop uses HDFS (Hadoop Distributed File System) file system**.** The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) . It is designed to run on large clusters (thousands of computers) of small computer machines in a reliable, fault-tolerant manner. It is based on master/slave architecture where master has a single NameNode which manages the metadata of filesystem and one or more slave DataNodes which is used to store actual data. Advantages of Hadoop are distributed storage, Security, Reliability, Speed, Efficiency, Availability and lots more.

Hive:

Apache Hive is database which is used for data analysis, storage, and summarization. It is build on top of hadoop. Hive provides an SQL-like interface for querying data stored in HDFS or different databases.

Flume:

Apache Flume is a flexible tool used to collect, stream, aggregate large amounts of data in to HDFS. It is a distributed service. It is used to sink Twitter's data into HDFS. It is reliable, robust and fault tolerant. It provides mechanism for failover and recovery.

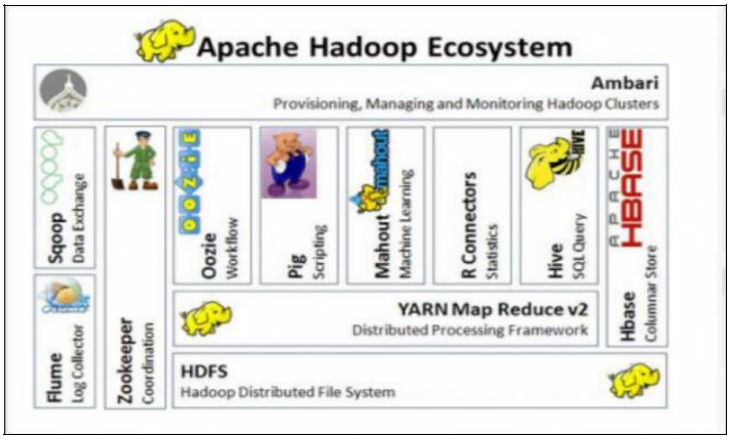

Figure 1. Hadoop Ecosystem

 ISSN: 2454-4248
576 – 578

III. ALGORITHMS

1) Naïve Bayes:

Naïve Bayes is a classification technique based on Bayes theorem with an assumption of independence between predictors. It calculates conditional probability of positive, Negative and Neutral tweets and classifies accordingly.

2) Decision Tree Algorithm:

The ID3 algorithm of decision tree, with 'Information Gain' function as the feature selection parameter was used to classify the tweets.

3) K-Nearest Neighbors:

A k-NN predicts the class of a tuple based on the class of the k nearest neighbors. K was chosen as 3, which is usually a common value used, and the Euclidean distance measure was used. The majority nearest tuple decided the class of the new tweets.

4) Support Vector Machine:

In SVM, each data item is plotted as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that differentiate the two classes(positive, Negative) very well.

Following is the result table obtained for the different approaches:

Table 1. Comparison of algorithms.

Algorithm	Accuracy	Precision
KNN	0.667	0.750
Naive Bayes	0.764	0.854
Decision Tree	0.654	0.732
Kernel SVM	0.561	0.632

The Naïve Bayes led to the highest accuracy while the Machine Learning model of KNN followed with the second highest accuracy.

5) Porter Stemmer Algorithm:

Stemming : It is process to convert a word to its root word. It is one of the phase in preprocessing the tweets.

e.g. Running – Run

Porter Stemmer Algorithm is most popular and based on suffixes in English language. It is an example of Heuristic Method. It is based on different set of rules like:

ATIONAL → ATE(relational → relate).

The rules are of the form:

(condition) S1 -> S2 Where S1 and S2 are suffixes Conditions:

Table 2. Porter Stemmer Rules

M	The measure of the stem
*S	The stem ends with S
v	The stem contains a vowel
*d	The stem ends with a double consonant
*o	The stem ends in CVC (second C not W, X, or Y)

.

VI. SYSTEM DESIGN

Twitter's API is used to collect data. Sink data into HDFS using Flume. For processing tweets 3 steps are involved tokenizing, stemming and remove stop words and then feature vector is extracted. After that HIVE can be used for twitter posts analysis. Then Mapreduced programs are used to classify tweets , It will also give count of people for positive, negative, neutral tweets..And finally display percentage in the form of piechart.

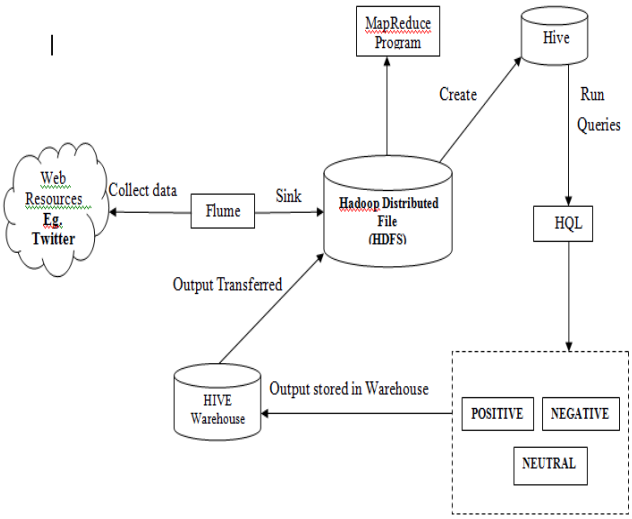

Figure 2. Architecture Diagram.

 ISSN: 2454-4248
576 – 578

VII. IMPLEMENTATION

1. Installation of Hadoop using Cloudera

Cloudera is a user friendly tool for installation of hadoop on your machine.It also provides graphical interface to user for installing and managing hadoop components like flume, hive, hue, oziee, pig, etc.

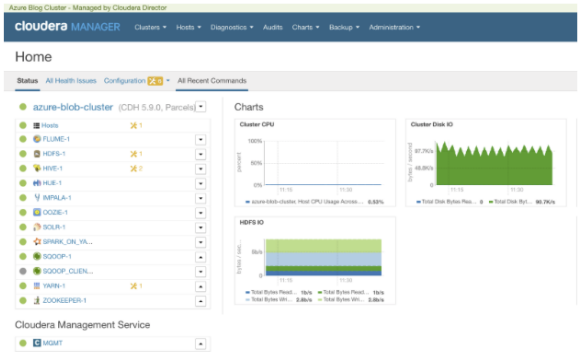

Figure 3. cloudera Dashboard

2. Accessing Twitter Data and storing it into HDFS using Flume Tool

For accessing the data log in to twitter. After that create a new application on twitter.After agreement you will get Consumer Key, Consumer Secret, Owner Key and Owner Secret ID for accessing data. After the creation of access token you will be able to get tweets. Add flume service using cloudera. Download the flume-sources-1.0-SNAPSHOT.jar and add it to the flume classpath in the conf/flume-env.sh file.Set Flume Classpath. Edit flume.conf file as follows and set different paramaters like consumer key, secret key etc:
Set Flume Classpath.
Edit flume.conf file as follows:

```
TwitterAgent.channels = MemChannel
TwitterAgent.sources = Twitter
TwitterAgent.sinks = HDFS
TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = <required>
TwitterAgent.sources.Twitter.consumerSecret = <required>
TwitterAgent.sources.Twitter.accessToken = <required>
TwitterAgent.sources.Twitter.accessTokenSecret = <required>
TwitterAgent.sources.Twitter.keywords = hadoop
TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://hadoop1:8020/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
TwitterAgent.channels.MemChannel.type=memory
TwitterAgent.channels.MemChannel.capacity=10000
```

Figure 4. Flume.conf file

3. Configuration of HIVE Build the JSON SerDe

1. Add jar to hive.
2. Create table.
3. Use Mapreduce programs for analysis.
4. Visualization : use python for creating dashboard(pie-chart).

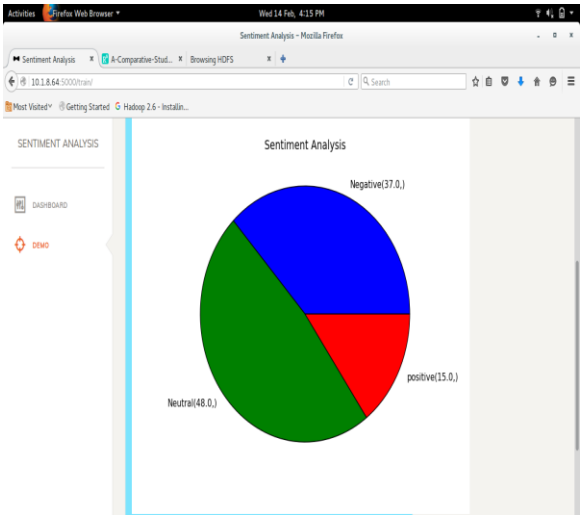

Figure 5. System Dashboard

VIII. CONCLUSION

In our research we analyzed sentiment of twitter data using k-NN, Naive Bayes, Kernel SVM, Decision Tree, Random Forest, and ANN. Along with this we found which metrics are most influential in forming a summary. The best performing models turned out to be Naïve Bayes.

We hope that our results benefit the future development of this project in terms of the selection of model and thus increasing accuracy.

IX. REFERENCES

[1]. International Journal of Engineering Research and General Science Volume 3, Issue 6, November-December, 2015ISSN 2091-2730

[2]. Jayashri Khairnar1, Mayura Kinikar2" Sentiment Analysis Based Mining and Summarizing Using SVM-MapReduce" International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, pp 4081-4085.

[3]. Imperial Journal of Interdisciplinary Research (IJIR)Vol-3, Issue-2, 2017 ISSN: 2454-1362, http://www.onlinejournal.in

[4]. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning. Proceedings of EMNLP.

[5]. Shalom Mathews et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (4) , 2016, 1955-1959

[6]. Manish Wankhede et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (6) , 2016, 2402-2404

[7]. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Web Site: www.ijettcs.org

[8]. Amit Gupte et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6261-6264.