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 5, May 2016

A Survey on Twitter Sentiment Analysis Using Hadoop

Anita Kumari[1], Anjali Kante[2], Shriya Samak[3], Ajinkya Ingle[4,] J.M.Kanase[5]

Student, Dept. of Computer Engineering, PES's Modern College of Engineering, Pune University, Pune, Maharashtra, India[1]

Student, Dept. of Computer Engineering, PES's Modern College of Engineering, Pune University, Pune, Maharashtra, India[2]

Student, Dept. of Computer Engineering, PES's Modern College of Engineering , Pune University, Pune, Maharashtra, India[3]

Student, Dept .of Computer Engineering ,PES's Modern College of Engineering, Pune University , Pune, Maharashtra, India[4]

Assistant Professor, Dept. of Computer Engineering, PES's Modern College of Engineering, Pune, Maharashtra, India[5]

ABSTRACT: In todays highly developed world, every minute, people around the globe express themselves via various platforms on the Web. And in each minute, a huge amount of unstructured data is generated. This data is in the form of text which is gathered from forums and social media websites. Such data is termed as big data. User opinions are related to a wide range of topics like politics, latest gadgets and products. These opinions can be mined using various technologies and are of utmost importance to make predictions or for one-to-one consumer marketing since they directly convey the viewpoint of the masses. Here we proposed to develop a Sentiment Analysis Application to analyze the sentiments of Twitter users through their tweets in order to extract what they think

Here we have used dictionary based approach for sentiment analysis for which we have implemented MapReduce algorithm which executes mapper and reducer class. In mapper class we compare each token with positive and negative dictionaries and based on that it assigns values to the token, then reducer class calculates sum of positive and negative sentiments.At the end result is displayed in the form of bar graphs and pie charts

KEYWORDS: Opinion mining; MapReduce; Sentiment Analysis; Dictionary based approach; Hadoop ;Cluster; Unstructured data.

I. INTRODUCTION

Sentiment Analysis:

Sentiment analysis also known as opinion mining.The process of computationally identifying and category opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product Sentiment Analysis is the process of detecting the contextual polarity of text. In other words,

it determines whether a piece of writing is positive, negative or neutral. Hence, there is a need to develop a product which can analyze opinions of people.This product will be useful in increasing market value of industries.As well as satisfy needs of customers.

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 5, May 2016

Motivation of the project

Today we are living in the world which is surrounded by 99 percent of data. There are different microblogging sites where users express their views about different products these views are nothing but opinions of people and it will go waste if it is not used in proper way so there is a need to use opinions of people in improving productivity, usefulness, functionality of particular product or application or technique or any entertainment resource.Hence, there is a need to develop a product which can analyze opinions of people. This product will be useful in increasing market value of industries. As well as satisfy needs of customers.

II. RELATED WORK

TITLE	AUTHOR	DESCRIPTION
Sentiment Analysis For Movie Review.	Shravan Vishwanathan	In this paper they have analyzed sentiments of people to predict the result that are based on users' opinion about movies.
Sentiment Analysis of Twitter Data	Apoorv Agarwal BoyiXie Ilia Vovsha Owen Rambow Rebecca Passonneau	In this paper they have used two methods:- (1) Introduction of POS-specific prior polarity features. (2)Use of a tree kernel to Obviate the need for tedious feature engineering.
Real Time Sentiment Analysis of Twitter Data Using Hadoop	Sunil B. Mane Yashwant Sawant Saif Kazi Vaibhav Shinde	This paper describes real time sentiment analysis of twitter.
Evaluation Datasets for Twitter Sentiment Analysis	Hassan Saif, Miriam Fernandez, Yulan He and Harith Alani	In this paper they have presented an overview of eight publicly available and manually annotated evaluation datasets for Twitter sentiment analysis
Twitter sentiment analysis Using apache storm	Ishana Raina Sourabh Gujar Parth Shah, Aishwarya Desai Prof. B.K.Bodkhe	Here they have proposed a system to analyze the tweets of Twitter users through their tweets in order to extract what they think. We classify their sentiments into three different polarities – "positive", "negative" and "neutral
Decision Making Using Sentiment Analysis from Twitter	M.Vasuki J.Arthi K.Kayalvizhi	This paper focused to predict the polarity of words and then classify them into positive and negative feelings with the aim of identifying attitude and opinions that are expressed in any form or language.
Hive – A Petabyte Scale Data Warehouse Using Hadoop	Ashish Thusoo Joydeep Sen Sarma Namit Jain Zheng Shao Prasad Chakka Ning Zhang	This paper describes sentiment analysis of data using Hive data warehouse.

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

International Journal of Innovative Research in Computer and Communication Engineering

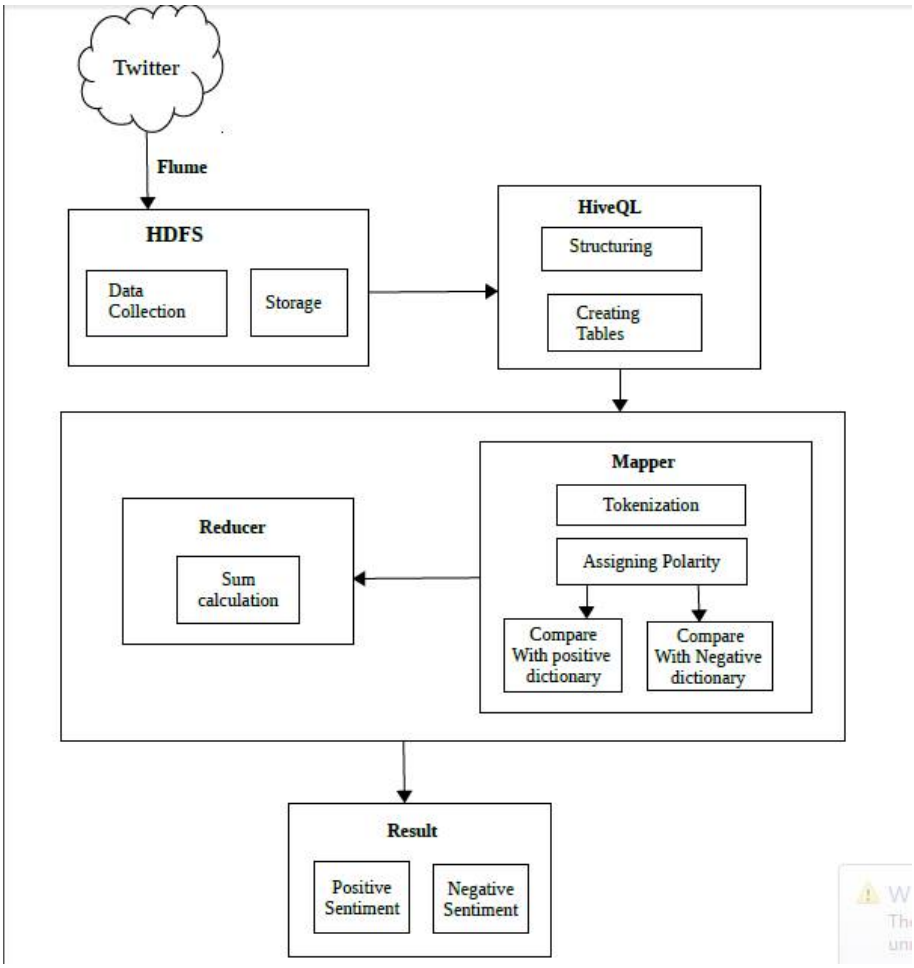
(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

III. PROPOSED ALGORITHM

A. *Design Considerations:*

Architectural design:



1. Twitter :

Twitter is an online social networking site which enables people to share their opinions about any trending topic in the form of short messages which are called tweets. Twitter datasets are freely available and it can be used to extract different sentiments of people about any topic or product and its beneficial to different companies for increasing their market value.

2. Flume : Apache Flume is a distributed, reliable, and available service for efficiently col-lecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS).

3. Data collection : Data sets are collected from twitter using following steps:

- Creating Twitter application: At first we are creating twitter application to get access to Consumer key,Consumer secret key, Access token key and Access token secret key.These keys are essential to collect data from twitter application.
- Injection of data to HDFS: For downloading datasets we passed keyword of related topic into config file of flume and we also added different token keys into the config file then we run flume agent from command line using command as soon as flume gets configured correctly it starts downloading tweets based on keyword and it injects tweets into 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 5, May 2016

4. Hive : Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis.

- Structuring: Datasets we have downloaded from twitter is in unstructured format which cannot be analyzed so we have used hive queries to convert unstructured data into structured data
- Creating tables: From unstructured data we are selecting some fields that is twitterId ,message and locations for creating different tables for analysis and this is accomplished by using HiveQL

5. Mapper :

- Tokenization: For analysing each tweet we need to break it into individual words which are called tokens and these tokens are compared with dictionary in order to generate sentiments.
- Assigning Polarity: we have created positive dictionary which contains positive words and negative dictionary which is collection of negative words. that we have generated are compared with positive dictionary if it is a positive word and it is compared with negative dictionary if it is a negative word. Based on positive and negative word polarities are assigned to the tokens.

6.Reducer : Sum Calculation:- It calculates sum of positive and negative words and result is displayed.

B. *Description of the Proposed Algorithm:*

After collecting twitter data we are preprocessing it using using Hive which includes structuring and tokenization. In tokenization we are breaking each tweet into individual tokens and these tokens are stored in mdata array which is of type string.

```
Map ( key , va lue , c o n t e x t )
        1 . f o r i from 1 t o mdata . l e n g t h
                a . If key i s p o s i t i v e
                        1 . compare key wi t h p o s i t i v e d i c t i o n a r y
                        2 . As s ign v a l u e 1
                b . e l s e I f key i s n e g a t i v e
                        1 . Comapare key wi t h n e g a t i v e d i c t i o n a r y
                        2 . As s ign v a l u e □1
                c . e l s e a s s i g n v a l u e 0
                d . End i f
        2 . End for
```

This function implements mapping in which it compares each token from mdata with positive and negative dictionary and assigns value accordingly

```
Reduce ( key , va lue , c o n t e x t )

        1 . Ca l c u l a t e sum of p o s i t i v e words
        2 . Ca l c u l a t e sum of n e g a t i v e words
```

IV. PSEUDO CODE

```
void map(LongWritable key, Text value, Context context)
{
for(int i=1;i<mdata.length;i++)
{
  if (dictpos.containsKey(mdata[i]))
{
        outputkey.set(mdata[0]);
        context.write(outputkey, one);
```

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 5, May 2016

```
} else if (dictneg.containsKey(mdata[i]))
{
            outputkey.set(mdata[0]);
            context.write(outputkey, minusone);

}
else {

             outputkey.set(mdata[0]);
             context.write(outputkey, zero);
}

}
}
```

V. CONCLUSION AND SCOPE

This project will give us hands on experience of handling and parallel processing of huge amount of data. Data collection process will introduce us to Java twitter streaming API. We will get exposure to work with prominent parallel data processing tool: Hadoop.

Apache Hadoop framework is gaining significant momentum from both industry and academia as the volume of data to analyze growth rapidly. This project will help us not only to gain knowledge about installation and configuration of hadoop distributed file system but also map reduce programming model. Amongst the many fields of analysis, there is one field where humans have dominated the machines more than any the ability to analyze sentiment, or sentiment analysis.
The future of this data analysis field is vast. This project not only analyses the sentiments of the user but also computes other results like the user with maximum friends/followers, top tweets etc. hence hadoop can also be effectively used to compute such results in order to determine the current trends with respect to particular topics. This can be very useful in the marketing sector.

Scope:

- Business Analytics: Consumers opinions provide valuable information about the companies as they help to understand how their products and services are perceived. So sentiment analysis is used in: Consumer voice, Brand reputation of the products, Online advertising: Blogger Centric Contextual Advertising and Dissatisfaction oriented online advertising, On-line commerce
- Politics: Sentiment analysis is used in voting advise applications and clarification of politicianspositions
- Public Actions: Sentiment analysis gives an important contribution in monitoring real world events for example for monitoring critical information about earthquake locations and magnitude, riot locations, this monitoring helps policy makers to minimise damage in areas which are expected to be affected next by such events.
- Policy or government-regulation proposals: Another important application of sentiment analysis is the monitoring of the opinions that people submit about pending policy or government regulation proposals
- Intelligent transportation system: A new emerging domain of sentiment analysis is Intelligent transportation system(ITSs), for the completeness of ITS space, it is necessary to collect and analyze the public opinions exchange. Traffic sentiment analysis has been developed which allows analysing the traffic problem in a humanizer way

REFERENCES

[1] Changbo Wang, Zhao Xiao, Yuhua Liu, Yanru Xu, Aoying Zhou, and Kang Zhang, "SentiView: Sentiment Analysis and Visualization for Internet Popular Topics", IEEE Transactions On Human-Machine Systems, Vol. 43, No. 6, November 2013.

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 5, May 2016

[2] Efthymios Kouloumpis, Theresa Wilson, Johanna Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!", Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media Rushabh Mehta, Dhaval Mehta, Disha Chheda, Charmi Shah and Pramila M. Chawan, "Sentiment Analysis and Influence Tracking using Twitter" in International Journal of Advanced Research in Computer Science and Electronics Engineering, Vol 1, Isuue 2, May2012.
[3] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau, "Sentiment Analysis of Twitter Data", Department of Computer Science, Columbia University.
[4] Bo Pang and Lillian Lee, "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008) Aditya Pal & Scott Counts, "Identifying Topical Authorities in Microblogs", WSDM'11, February 9–12, 2011, Hong Kong, China, Copyright 2011 ACM.
[5] Jianshu Weng, Ee-Peng Lim, Jing Jiang, Qi He, "TwitterRank: Finding Topic-sensitive Influential Twitterers", WSDM'10, February 4–6, 2010, New York City, New York, USA Copyright 2010 ACM.

BIOGRAPHY

Anita Kumari is a final year student in the Computer Engineering Department, PES's Modern College of Engineering, Pune University. Currently, she is doing his BE project on " Twitter Sentiment Analysis using Hadoop". Her area of interests are Big Data And Analytics,Databases, Networking and Information Security.

Anjali Kante is a final year student in the Computer Engineering Department, PES's Modern College of Engineering ,Pune, Pune University. Currently, she is doing his BE project on "Twitter Sentiment Analysis using Hadoop".Her area of interests include big data and analytics,database and networking

Shriya Samak is a final year student in the Computer Engineering Department, PES's Modern College of Engineering Pune, Pune University. Currently, she is doing her BE project on "Twitter Sentiment analysis using Hadopp". Her area of interests include databases,Hadoop and big data

Ajinkya ingle is a final year student in the Computer Engineering Department, PES's Modern College of Engineering Pune,Pune University.Currently he is doing his BE project on "Twitter Sentiment Analysis using Hadoop". His area of interests include DBMS, big data and hadoop

J.M.Kanase is an assistant professor in the Computer Engineering Department, PES's Modern College of Engineering Pune, Pune University. She is an expert in Distributed computing system and operating system.She has more than 5 years of experience in these fields.