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. 3, Issue 7, July 2015

A Big Data Methodology for Sentiment Analysis of Twitter Data

Supraja.G.S[1], Dr Jharna Majumdar[2], Shilpa Ankalaki[3]

Dept of CSE, Nitte Meenakshi Institute of Technology and Management, Bangalore, India[1]

Professor and HOD, Dept of CSE, Nitte Meenakshi Institute of Technology and Management, Bangalore, India[2]

Asst. Prof, Dept of CSE, Nitte Meenakshi Institute of Technology and Management, Bangalore, India[3]

ABSTRACT: Now a day there is drastic increase in the usage of internet among masses. The usage of internet is leading to the generation of large data sets. Efficient handling of such large data (also known as Big Data) is an ongoing important research across the world. Handling of Big Data includes storage and processing of Big Data. Also analysing the Big Data, this includes discovering the pattern or knowledge discovery from the Big Data which is called as "Big Data analysis". Increase in Internet usage is mainly because of the social media popularity. Twitter is an online social networking site which allows users to post real time short messages. These messages are limited to 140 characters and called as "tweets".
In this paper we are proposing a methodology to collect and store live twitter data and perform sentimental analysis using machine learning techniques and provide some prediction. To store the live data fetched we are using MongoDB a NoSQL database, the output of the analysis will be trend analysis with different sections that is positive, negative and neutral.

KEYWORDS*:* BigData, Sentimental analysis, Machine learning, MongoDB, Twitter

I.INTRODUCTION

1.1 Big data

According to McKinsey Report [1] Big Data can be defined as "data sets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" Those datasets are generated mainly through the internet usage, mobile devices, sensor networks, enterprise system and organization. Big data is not only about volume it also consider variety and velocity. Big data generated can be structured, unstructured and semi structured.

[2]With modern technologies such as devices, machinery, vehicles embedded with sensors and increase of internet usage among masses especially "social media such as Facebook and also GPS devices" generating large amount of unstructured data which is complex. This is called as Big Data. It can be effectively utilize by combining with structured Data from traditional business applications such as "Customer Relationship Management or Enterprise resource Planning" the advancement of technology has made collecting and processing of Big Data easier and in the mere future Big Data will definitely grow enormously and in the mere future

Internet usage among masses is increasing drastically mainly because of popularity of social media. Various Big data technologies like schema-less databases or NoSQL databases, Hadoop, Hive, Pig, PLATFORA aims at collection, processing and storing of big data in a cheaper and effective way. Big data does not only mean about storing the large amount of data but also storing and analysing of data and predicting some pattern or trend in business intelligence.

1.2 Sentimental Analysis

[3] "Sentiment analysis can be defined as the analysis by treating computationally user or public opinion or sentiment and also subjectivity in an text especially text obtained from social media". It can be use to know the opinion of user with respect to the particular topic. Some cases it can be used for judgement like product success or failure in impressing the crowd when the new or updated version will be released to the market.

[4]Twitter is an online social networking site which allows users to post real time short messages which is limited to 140-character those short messages are called "tweets". Registered users have access to read and post tweets, but

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. 3, Issue 7, July 2015

unregistered users can only read the tweets. Sentiment analysis also known as opinion mining can be defined as analysis of textual data particularly social media data.

1.3 MongoDB

[5]MongoDB is document database developed by 10gen in 2008. It is a GPL open-source and written in C++ and its query language is JavaScript, it provides interactive JavaScript shell for database management system. MongoDB supports its own ad-hoc query language.

MongoDB has both strengths like it is easy to install, PHP module is available, replication is easy including master-master support and weaknesses like it does not ACID transactions and Joins , but overall the performance is quite good and it doesn't have as many restrictions and limitations as other NoSQL databases.

1.4 Machine learning

[6]"Machine learning-stream of artificial intelligence, which focus on research to train the system to identify patterns and make decision based on the available data".

Machine learning contains methods that can train the system to predict future data from available data and also to take decision on how to collect more data and to identify patterns from the available data automatically.

There are two categories in machine learning, the former one is supervised learning and the later one is unsupervised learning technique. In predictive or supervised learning approach the aim is to learn a mapping from inputs.

x →outputs y,

Given, set of input-output pairs

$$D = \{(x_i, y_i)\}_{i=1}^{N}$$

Where D is defined as training set, and N is designed as number of training examples.

II. METHODOLOGY

Twitter one of the popular social media has gained popularity dues to availability of valuable information for governments, business across the world. Potential information is available from Twitter will be textual data which can be easily accessed and cleaned. The textual data can be used for sentimental analysis of the public towards particular product or particular person. Also, it can be used for trend analysis, cluster analysis or classification analysis.

Research in sentimental analysis involves selecting appropriate clustering techniques from the implemented clustering techniques based on execution time analysis.

2.1 Proposed Methodology

The objective of this research is to develop a methodology to perform sentimental analysis of live twitter data and categorize the tweets in to positive opinion, negative opinion and neutral opinion. Also choose the best clustering method based on execution time analysis. The steps followed in sentimental analysis of twitter data is as follows:

Step 1: Create Application and retrieve live data from Twitter

a. Create a Twitter Dev App
b. Authenticate the App and generate consumer and authentication keys
c. Provide Permission like Read or Read/Write to access the App
d. Implement the code to connect to Twitter App and get the required tweets for the Analysis based on some specific topic.

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. 3, Issue 7, July 2015

Step 2: Store the Tweets

a. Configure Mongo DB (It is document based database that is used to store big data. It Stores the data into JSON format)
b. Implement the code to connect to Mongo DB server.
c. Implement the code to connect to twitter app from mongo dB to get and store and process the tweets in JSON format.

Step 3: Pre-processing

This step involves remove the unwanted urls, words and symbols from the tweets and get the data points to generate the cluster as an input to the algorithms.

Step 4: Create the Clusters using machine learning algorithms

Three machine learning algorithms (K-mean [7], CURE [8] and BIRCH [9]) were used to cluster the data. These clusters will be input for the sentiment analysis step.

Step 5: Comparing the Effectiveness of Algorithms

Comparing the time taken to create large number of clusters by these machine learning algorithms.
This will provide the effectiveness of algorithms.

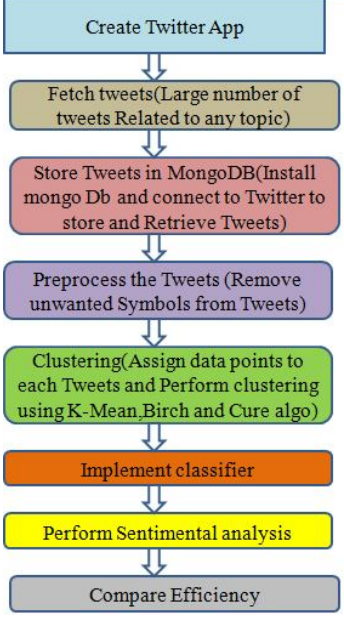
Step 6: Create Classifier

Classifier is created to use with different clusters to do the sentiment analysis. This will provide the base of sentiment analysis

Step 7: Create Classifier

"Sentiment analysis is the process of analysing the attitude of a user or a person with respect to some topic or attitude of all the context of a document".
I have used the custom classifier to determine the sentiment of each sentence for a particular topic. Input for these steps will be live tweets

Figure 1: Proposed Methodology



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. 3, Issue 7, July 2015

III.RESULTS AND ANALYSIS

3.1 Results of Proposed Sentimental Analysis System

Sample Tweet and Semantic Orientation

1: Negative Sentiment 2: Neutral Sentiment 3: Positive Sentiment

1	Kejriwal hits back at Modi says people are not lucky Hiting back at Prime Minister Narendra Modi for his
1	Modi is a tolerant forgiving man like most Hindus are But our enemies take ad
1	Fascism of RSS Modi s silence on the isue is a Sin Church vandalised in Delhi s Vasant Kunj area prayer it

Table 1 Negative sentiment

2	Ana shed light on BJP MODI today in the interview That s enough to embaras him
2	And if you mised hearing the Man Ki Bat episode you can listen to it here
2	Must WatchModi Obama Man ki Bat episode in e-bok format

Table 2 Neutral sentiment

3	Foreign Secretary S Jaishankar One of Modi s best decisions as prime minister
3	India can move from natural to best partner of US if it sticks to religious tolerance says Obama What say Modi
3	Modi is the best prime minister any one can get he is a brand

Table 3 Positive sentiment

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. 3, Issue 7, July 2015

3.2 Comparison of time taken by all three algorithms

Algorithm	Number of data points(tweets)	Time taken to create clusters (In seconds)
KMean	1000	1.022
KMean	10000	2.325
KMean	100000	5.607
Birch	1000	1.192
Birch	10000	2.133
Birch	100000	4.309
Cure	1000	1.678
Cure	10000	2.928
Cure	100000	3.257

Table 4 Comparision of efficiencies of algorithm

From above comparison table we can see that Kmean and Birch algorithms take almost same time to perform sentiment analysis for small number of tweets. But for large number of tweets Cure algorithm performs better than Kmean and Birch algorithms.

IV.CONCLUSION AND FUTURE WORK

In our project we have implemented and shown the results for three different clustering algorithms for sentiment analysis. All three algorithms provide an effective time to perform sentiment analysis for different and large number of tweets.
Kmean algorithm is performed for 2 and 3 clusters. In future it can be implemented for large number clusters. Cure algorithm also can be used for different scenarios for search and perform different clustering. Birch algorithm can be used more efficiently to perform clustering and give better result for sentiment analysis.
In future sentiment analysis can be optimized using different techniques for processing big data like Hadoop, elastic search and r analysis.
In future this system can be used with different social sites to perform sentiment analysis on large set of data sets and big data.

REFERENCES

[1] Internet source, aisel.aisnet.org
[2] Internet Source, www.victa.nl
[3]Internet Source, www.nowpublishers.com
[4] Akshi Kumar and Teeja Mary Sebastian, "Sentiment Analysis on Twitter" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012ISSN (Online): 1694-0814
[5] E. Dede, M. Govindaraju, D. Gunter, R. Canon, L. Ramakrishnan, "Performance Evaluation of a MongoDB and Hadoop Platform for Scientific Data Analysis" Proceedings of the 4th ACM workshop on Scientific cloud computing – Science Cloud13, 2013.
[6] www.cs.ubc.ca Internet Source
[7]ZHEXUE HUANG ,"Extensions to the *k*-Means Algorithm for Clustering Large Data Sets with Categorical Values"
[8]Sudipto Gulba, Rajeev Rastogi, Kyuseok Shim,"CURE: An Efficient Clustering Algorithm for Large Databases"
[9]Tian Zhang, Raghu Ramakrishnan, Miron Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases" SIGMOD '96 6/96 Montreal, Canada IQ 1996 ACM 0-89791 -794-4/96/00

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. 3, Issue 7, July 2015

BIOGRAPHY

Supraja G S Studying final year Mtech in Nitte Meenakshi Institute of Technology,Bangalore.She received Bachelors degree in 2012 from Reva institute of technology and Management,Bangalore . Has Industry experience working on javascript technologies extjs,D3,Raphael. Area of interest includes Big Data and Data mining Techniques.

Dr. Jharna Majumdar is currently the Dean R & D and Professor and Head Dept of Computer Science and Engineering (PG), NMIT Bangalore. She served Defence Research and Development Organization, Govt. of India from 1990 to 2007 as Scientist G and Head of Aerial Image Exploitation Division, Aeronautical Development Establishment, Bangalore. Dr. Majumdar worked as a Research Scientist in the area of 'Robotics and Automation' at the Institute of Real Time Computer Systems and Robotics, Germany during 1983 to 1989 and at the Stanford Research International California USA 2002.
Dr. Majumdar published about 110 reviewed technical papers in national/international journals and conference proceedings. She received a large number of awards for her work from DRDO. Some of her award includes the award received from President, Stanford Research International California USA, Performance Excellence Award from Prime Minister of India, Dr V M Ghatage award, Dr. Suman Sharma Award, Dr. Kalpana Chawla Memorial Award etc.
Her project with ISRO and a consortium of 7 Engineering Colleges with NMIT as the lead Centre for building the first PICO satellite in India took off the orbit on mid May 2010. Nitte Amateur Satellite Tracking Centre (NASTRAC) developed by a team of students from NMIT under her guidance is a patent and the first Tracking Station of Small Satellites developed in the country. The research team of robotics under her guidance has developed the first Robotic Exhibit at the Birla Science Centre Hyderabad on Jan 2012.
Her current research areas include Robotics, Image and Video Processing, Pattern Recognition, AI, Computer Vision in Embedded Processor for Aerial and Ground Based Platforms and Big Data.

Shilpa Ankalaki is currently working as asst.prof in Nitte Meenakshi Institute of Technology and Management. She completed her M.Tech from Nitte Meenakshi Institute of Technology and Management,Bangalore. Her Area of Interest includes Digital Image Processing, Data Mining Technology,Pattern Recognition.

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. 3, Issue 7, July 2015