OPINION ANALYSIS OF SOCIETAL ISSUE – "THE JALLIKATU PROTEST" WITH TWITTER DATA USING BIG DATA ANALYTICS

C. Premalatha
premalatha.chinnasamy@srec.ac.in
Assistant Professor, Department of Information Technology,
Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu, India

ABSTRACT

Data are scalable, sceptical, versatile and diverse compared to good old structured one that is derived from specific sources. In addition, the storage, handling and analysis of such a trendy data results in a very complicated containment, processing and visualization problems. To overcome these disadvantages on trendy unstructured data there emerged a concept called Big Data Analytics. It uses numerous tools and techniques that solved the problem from data storage to visualization. Analytical sandbox is the most focal container that stores and handles data in a very appropriate manner. Big data tools provide a wider range of analytical and visualization techniques that results in efficient graphical view of modern data. This paper focused on analysis of twitter data in which text mining was inculcated at the start then proceeded with analytical tools to analyse and visualize the opinions of society for jallikattu issue. It produced a wordcloud that represents highly rated talks on that issue.

Keywords: Big data Analytics, Analytical Sandbox, Unstructred data, Big data tools, Analytical and Visualization techniques, twitter

1. INTRODUCTION

In an environment, the most focal thing that aids in decision making is data. It can either be historical or current, that makes up the analytical world a trending one. In spite of the importance given to data, the priority lies in its storage. Initially data was stored in the form of rows and columns for which a concept was coined as Database Management System. In latter, happens all the storage and extraction process which totally focused on structured data. Later on due to the evolution in data formats, that is inclusion of multimedia contents was quiet larger in all aspects, so data containment extended to data warehouse that stores semi structured data. In recent trends, due to increased usage of social sites having plentiful data that occurs in varieties like structural, semi-structural, Quasi-structural and unstructural, the storage trend produced a concept of Analytical Sand box containing all variety of data from which higher level of analysis and visualization can be done. The thrust for accessing such varieties of data shaped Big Data Analytics. It provides intelligible, scalable, diverse and defined approach for data analysis and visualization.

The process of mining intelligible data from a source is the most important and complicated task as it must provide quantitative, technical, communicative and coherent knowledge to the viewer. The process starts with collection, storage, data handling, analysis and visualization. These can be done through various algorithmic approaches and tools as well. The most challenging task is extracting and analysing real time social data that contains variety of data that has to be handled intelligibly. Big data analytics helps in solving such kind of problems more effectively.

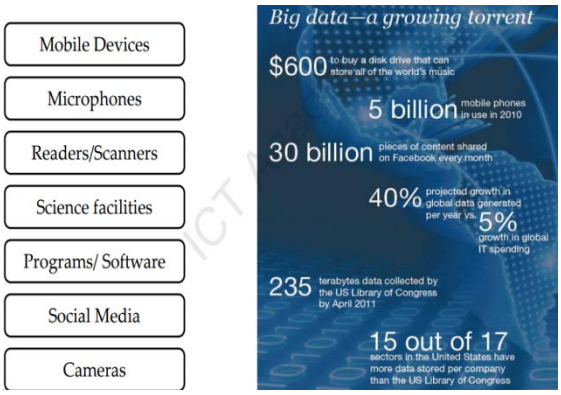

Fig.1.1. Data Generation Points

2. BIG DATA ANALYTICS

Big data is termed as extremely large dataset that is so large or complex to do any analysis to reveal patterns, trends and associations that can not be analysed with the help of traditional data processing applications. Big data is a field dedicated to the analysis, processing and storage of large collections of data that frequently originate from disparate sources. Big data solutions and practices are typically required when traditional data analysis, processing and storage technologies and techniques are insufficient. Specifically, Big data addresses distinct requirements such as combining of multiple unrelated datasets, processing of large amount of unstructured data and extracting of hidden information in a time sensitive manner. Big data lifecycle generally involves identifying, procuring, preparing and analyzing large amounts of raw, unstructured data to extract meaningful information that can serve as an input for identifying patterns, enriching existing data and performing large-scale searches.

Data Analytics [1] is a broader term that encompasses data analysis. Data analytics encompasses collection, organization, storing, analyzing and governing data. The term includes the development of analysis methods, scientific techniques and automated tools. In Big data environments, data analytics has developed methods that allow data analysis to occur through the use of highly scalable distributed technologies and frameworks that are capable of analyzing large volumes of data from different sources. Data analytics enable data-driven decision making with scientific backing so that decisions can be based on factual data and not simply on past experience or intuition alone.

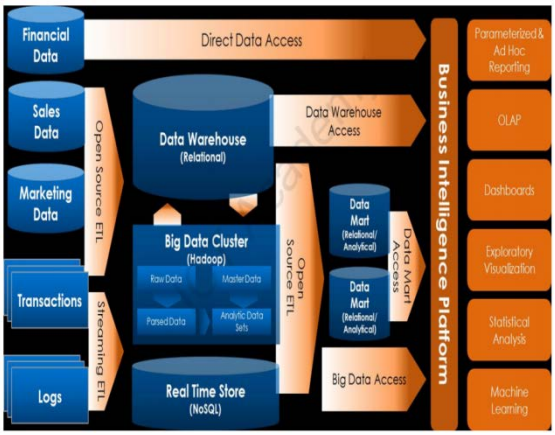

Fig.2.1. Big Data Architecture

For a dataset to be considered Big data, it must possess one or more characteristics [1] that require accommodation in the solution design and architecture of the analytic environment. This section explores the five Big Data characteristics that can be used to help differentiate data categorized as "Big" from other forms of data. They are

- Volume-Distinct data storage and processing
- Velocity-Data can arrive at fast speed and enormous datasets can accumulate within very short periods of time
- Variety-Multiple formats and types of data that need to be supported
- Veracity-Quality or fidelity of data
- Value-Usefulness of data

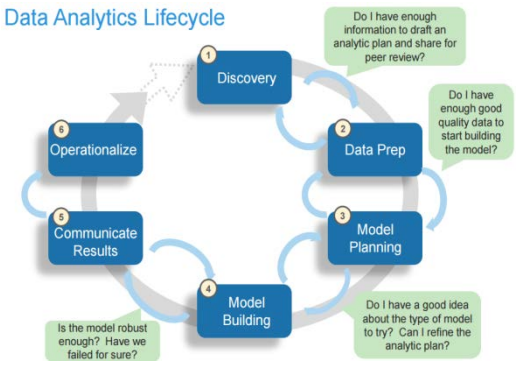

Fig.2.2. Data Analytics Life cycle

3. TYPES OF DATA

Data sources can be varied and it can be any format like audio, video, text, table, etc., such that it is categorized as follows:

3.1 Structured Data[1]

It conforms to a data model or schema and is often stored in tabular form. It if frequently

generated by enterprise applications and information systems like ERP and CRM systems. Due to the abudance of tools and databases that natively support structured data, it rarely requires special consideration in regards to processing or storage.

3.2 Semi-structured Data

It has a defined level of structure and consisitency, but is not relational in nature. Instead, it is hierarichal or graph-based. This kind of data is commonly stored in files that contain text. Due to textual nature of this data and its conformance to some level of structure, it is more easily processed.

3.3 Quasi-structured data

It is totally intuitive, emergent, pseudo, guess, apply a rule and refine process

3.4 Unstructured Data [1]

Data that does not conform to a data model or data schema. This form of data is either textual with various blog posts or it can be binary with files like image, audio, video data and often conveyed via files that are self-contained and non-relational. Special processing is usually required to process these dataand it cannot be directly processed using SQL. If it is required to be stored within relational database,it is stored in a table as a Binary Large Object(BLOB). Alternatively, a Not-only SQL(NoSQL) database is a non-relational database that can be used to store unstructured data alongside structured data.

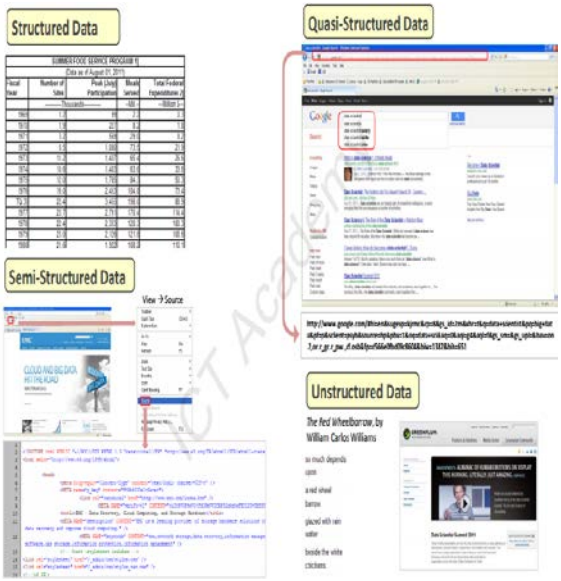

Fig.3.1. Types of Data

4. PACKAGES IN BIG DATA ANALYTICS

R Language[2] provides rich graphical facilities for data analysis and supports basic graphs to advanced graphs. R supports graphics to create basic charts like pie, bar and line. R also has extension of many packages for data visualization. Visualization of data helps to understand the trends in large datasets. R offers multiple packages for data analysis and provides interfacing with other tools for data analysis. There are more than 7769 packages available including Github, CRAN and Bioconductor. CRAN lists the view of basic primary packages in an organized way.

R package[2] is a collection of function, data and compiled code in a defined format and stored as library. Standard packages are installed in R for basic data management, analysis and graphical displays. Additional packages can be installed and loaded as required.

Table 4.1: List of Packages in R-tool

Data Visualization	ScatterPlot3D, BoxPlotDBL, GGPlot2, Giraph, AutoMap
Data Statistics	Banova, AdMit, FunChiSq, SPCOV
Data Transformation	Binr, Dummies, SM, SME, ACEPACK, Discretization
Classification	Caret, Class, AL3, AutoPLS, EasyNLs, ClogitL1
Clustering Basic	AKMeans, CKMeans, BayesClust
Regression	LmTest, NLSTools
ROC Analysis	ClustEval, GGROC

4.1 Packages for twitter

The list of packages used for accessing twitter is as follows:

- twitteR
- Rcurl
- base64enc
- httr
- tm
- wordcloud

4.2 Wordcloud

A word cloud is a technique for visualization of words typically associated with Internet keywords and text data. They are most commonly used to highlight popular or trending terms based on frequency of use and

prominence. A word cloud is a beautiful, informative image that communicates much in a single glance.

5. ANALYSIS OF TWITTER DATA

Analysing the opinion and attitude of a person through mining relevant data is termed as sentimental analysis or opinion mining. It focuses on the feel of person on particular topic in a forum or in a discussion. It is an emerging concept that has high echelon on positive and negative opinions. In this paper, a twitter data is considered for analysis. Using R, opinion analysis of jallikattu protest has been incorporated and the result is visualized using wordcloud. Based on the result, the opinion of the people in this particular issue has been visualized.

5.1 Analysis steps are as follows

- Install and import the packages such as twitter, RCurl, base64enc, httr, tm and wordcloud
- Include consumer_ key, consumer_ secret, access_token, access_secret for accessing the twitter data
- Setup twitter oauth
- Using search Twitter mine the texts related to jallikatu
- Preprocessing of extracted data is done using corpus() and tm_map().It includes following:
  - Remove punctuation
  - Content transformed
  - Remove stop word
  - Remove numbers
  - Strip off whitespaces
- Incorporate wordcloud() to have visualization effect of high rated individual

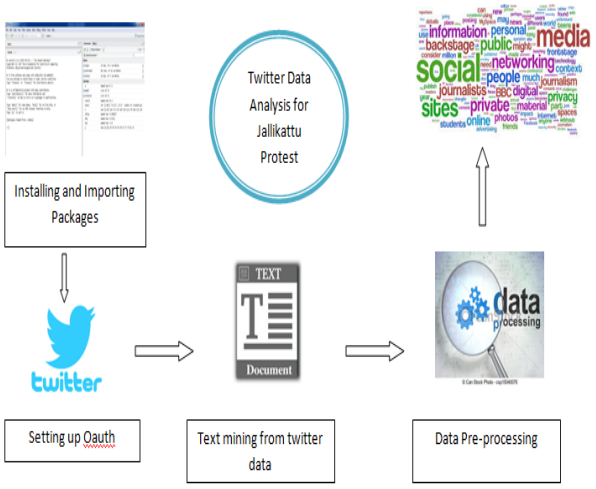

Fig.5.1.Opinion Analysis Process flow Model

5.2 Sample Analysis Report

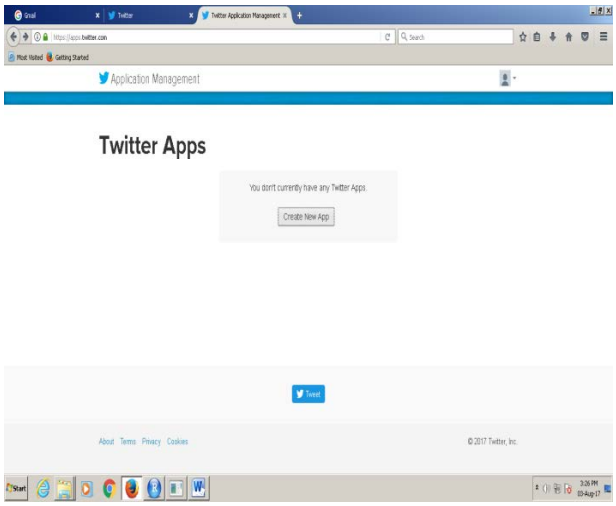

Fig.5.2.1. Twitter Application Management Site

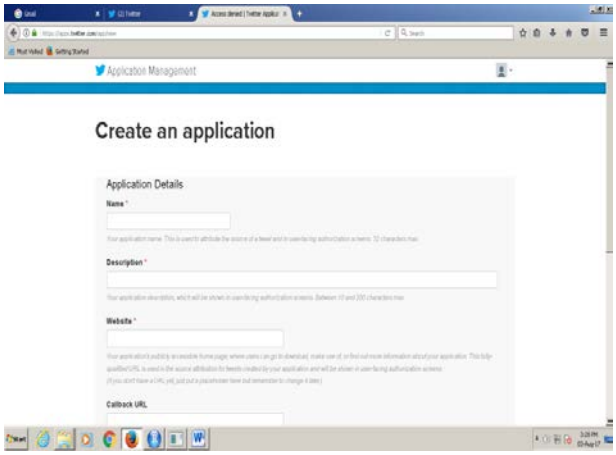

Fig.5.2.2. Twitter Application creation

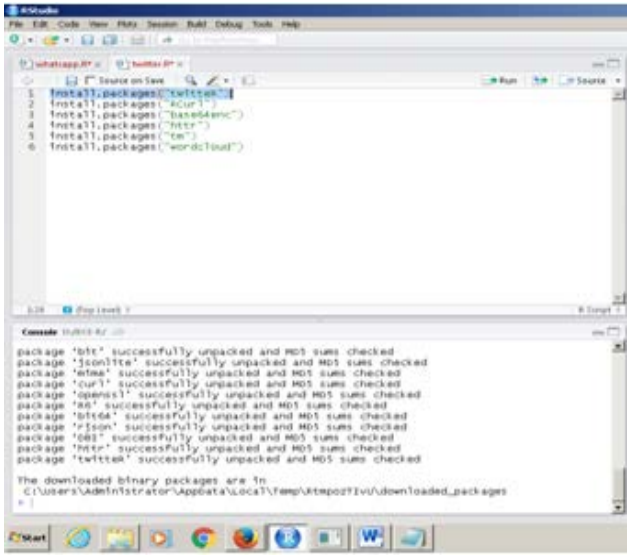

Fig.5.2.3. Importing required packages

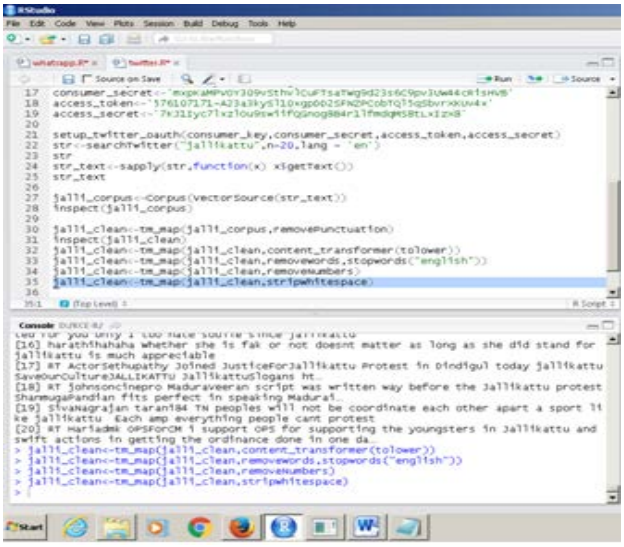

Fig.5.2.4 Pre-processing Twitter data after setting up oauth

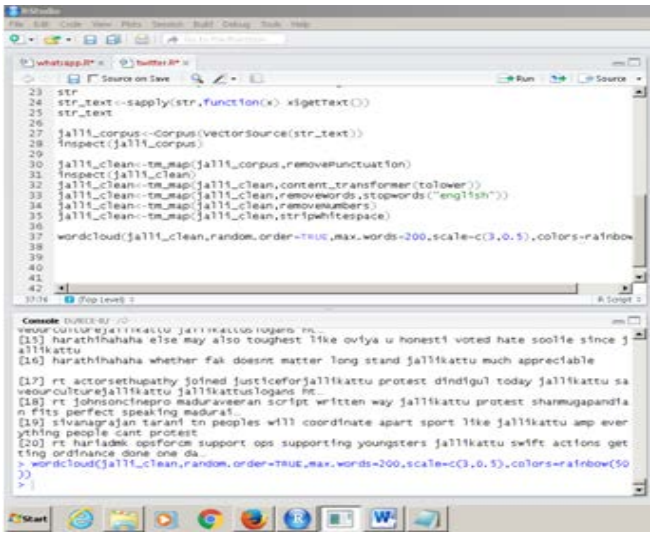

Fig.5.2.5. Data Visualization of Jallikatu protest opinions

Fig.5.2.6. Wordcloud of Jallikatu protest opinions

6. CONCLUSION AND FUTURE ENHANCEMENT

Opinions are the best review for any applications worldwide. Here, in this paper those opinions were analysed using emerging Big Data Analytical tool. Finally, it produced a result with different opinions of the people for the societal issues. Hence, maximum opinions are highlighted. This paper focused on the online data analysis. So opinion can be varied as well when time moves on such that this can be extended to complex online streaming analysis for other social networks

REFERNCES

[1]Thomas Erl, Wajid Khattak, Paul Buhler, “Big Data Fundamentals: Concepts, Drivers & Techniques”,Pearson India Education Service Pvt. Ltd. 2016

[2]V. Bhuvaneswari,“Data Analytics with R”, published by budca.in,ISBN:978-81-929131-2-4, Edition, 2016

[3]http://datascienceplus.com/sentiment-analysis-with-machine-leaning-in-r/

[4]http://sites.google.com/site/miningtwitter/questions/sentiment/sentiment

[5]http://andybromberg.com/sentiment-analysis

[6] Samiddha Mukherjee , Ravi Shaw, “ Big Data – Concepts, Applications, Challenges and Future Scope”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 2, February 2016

[7] Priyank Jain, Manasi Gyanchandani and Nilay Khare, “Big data privacy: a technological perspective and review”, Journal of Big Data, DOI 10.1186/s40537-016-0059-y

[8] Nitish Sinha, ―Using Big Data in Finance: Example of sentimentextraction from news articles‖; FEDS notes, March 2014

[9] Baker, Malcolm and Jeffrey Wurgler, 2007. "Investor Sentiment in the Stock Market", Journal of Economic Perspectives, vol. 21(2), pages 129-152