Twitter Sentiment Analysis in Healthcare using Hadoop and R

Vijay Shankar Gupta
Department of Computer Science
Birla Institute of Technology, Mesra
Ranchi, India
Email Id: vijayshgupta@gmail.com

Shruti Kohli
Department of Computer Science
Birla Institute of Technology, Mesra
Ranchi, India
Email Id: kohli.shruti@gmail.com

Abstract – Healthcare is a data-rich industry. Big data is fast, expanding rapidly and is highly varied. Twitter can serve as an important data source for providing real-time information that has stimulated companies in diverse domains to understand their consumers. In the healthcare field we will concentrate on patients, the illness they were suffering from, the hospital they preferred at the time of sickness, service provided to them and were they satisfied or not. The era of social media generates an enormous amount of data for analysis. The outcome information is beneficial for the health of individuals. Analyzed information is further used for the recommendation of hospitals at the time of ailment. Health care is still a nascent field undergoing rapid change and evolution due to continuous enhancements in medicine. It is based on the opinion and satisfaction rate of the patients. This work is conducted through an online survey and then based on the satisfaction rate, the hospitals are selected. The messages posted in Twitter called tweets are used for comparison of the sentiments of the patients that are correlated to the hospitals selected. The comparison shows the popularity of a hospital among patients at the time of ailment.

Keywords – Big Data, Hadoop, Healthcare, Hive, Flume, R, Sentiment Analysis, Twitter.

I. INTRODUCTION

In a spur of seconds, an enormous amount of data is generated, captured, and transferred through various media. Big Data is highly velocious, becoming more prominent and shortly will be gigantically large. Big data is alluring as it is influencing every aspect of life on earth. Big data is somewhat similar to planet developing a nervous system [1].

Big data is a term used by the IT industry to describe the voluminous amount of unstructured data an organization creates. We are now beginning to measure, analyze, visualize and respond to what is happening all over the world in real time as the species that we had never done before [1].

As we are moving to create a smart planet with smart cities, it is essential to have smart data ”Big Data” which can be explored for smart growth to benefit smart people. Big data is available in every technological sphere that one can imagine; thus we can see its presence in each and every field. Big data is ubiquitous and omnipresent. It is just phenomenal creating countless new digital puddles, lakes, tributaries and oceans of information [2].

The big data applications used in healthcare are boosted with a high level of the quality result, inexpensive operations, and quick optimization. This approach leads to the radical change in healthcare issues and research. Leveraging big data will certainly be part of the solution to controlling spiraling healthcare costs [3]. One thing is sure - the future will only bring more and more data to our doorstep. So, there is a need to analyze effectively and use all of that data with an aim to continuously improve the outcomes, processes, and operations.

II. BIG DATA: AN EMERGING ERA

A. Definition

Gartner is a world’s leading advisory organisation and research group. They explained big data as it is loaded with volume, velocity, and variety [4]. The information system which we are using is not made for the solution. It requires new forms of pre-processing and preparation to deal with this enhanced decision oriented insights, discoveries, and process optimization.

Manyika et. al.[5] had described big data as the next border for productivity and innovation. The big data provides not only an extensive prospective for the growth of individual companies. Mathiesen et al.[6],analyses the user behaviour by analysing the occurrence and co-occurrence frequency of keywords in user posts. Multinomial Nave Bayes is best for the segregation of tweets on the basis of different emotions.

One of the definitions given in The Digital Universe Study: Big Data technologies are a new generation of technologies and architectures. It is designed to extract economically from a

978-9-3805-4421-2/16/$31.00

massive data set of a wide variety of data by enabling high-velocity capture, discovery, and analysis. There are three key features of Big Data: the data itself, the analytics of the data, and the presentation of the results of the analytics [7].
In Horizon 2020, Big Data finds its place both in the Industrial Leadership and the Societal Challenges. The Big Data is a technique to extract, transform, analyze and visualize potentially massive datasets in a reasonable timeframe [8].

B. Related Works

Big Data is associated with three principal features i.e. volume, velocity and variety. But for healthcare domain we will consider 5Vs which are described in Table I [9].

TABLE I. BIG DATA DIMENSIONS IN HEALTHCARE

Volume	Large amount of data generated by organizations or individuals in terabytes or even Petabytes.
Velocity	Frequency and speed, at which data is generated, captured and shared.
Variety	Amalgamation of data from different sources like mobile, machine, social media.
Value	Exploiting the large amount of data to extract insight and draw valuable conclusions out of it.
Veracity	Results and outcomes of big data analytics are error free, credible and accurate.

Big data is made up of three kinds of data: Structured, Unstructured, and Semi-structured.
Structured data is an organized data in a predefined format. The formatted data resides in fixed fields within a record or file and has entities and attributes mapped. Structured data is used to query and report on predetermined data types [10].
Unstructured data is a set of data with a complex structure that might or might not have a repeating pattern [10]. Semi-structured data is also known as schema-less or self-describing structure. It refers to a form of structured data that contains tags elements to separate semantic elements. It also generates hierarchies of records and fields in the given data. Such type of data does not follow the proper structure of data models as in relation databases [10].
Five different streams or categories of information are comprised under Big Data [11,12]:

1. Social media and Web data: Streaming and interactive raw form of data from many social media websites such as Twitter, LinkedIn, Facebook, etc. It includes health-related blogs and websites like MedHelp. Data from Smartphone apps are in used for analyze customer behavior.
2. Data from Sensors: Readings coming from meters, sensors and other devices.
3. Huge transactional data: Medical billings and claims are available for research. These Health datasets are increasingly available in somehow unstructured or semi-structured formats.
4. Data from Biometrics: Dataset which includes Genetics, retinal scans, fingerprints and handwriting data. These biometric data would also include X-rays and other scans pictures slides, pulse readings and blood pressure.
5. Automated generated data: These are also unstructured or semi-structured data such as physician notes, paper documents, electronic medical records (EMRs) and email.

Opportunities for Big Data in Healthcare: Health Care Excellence and Efficiency: As of 2010, national health expenditures represented 17.9 percent of gross domestic product (GDP), up from 13.8 percent in 2000. At the same time, the frequency of chronic diseases like diabetes is rising and consuming a larger percentage of health care assets. Electronic Health Records (EHRs), coupled with novel analytics tools, unlock the door to mining information for the most valuable output of vast populations [11, 13].
Disease Detection: Sensors are progressively being used to monitor the key biochemical markers. The data coming from individual patients to HIPAA-compliant analysis systems are analyzed with real-time analytics. This analytics approach can alert particular patients to potentially adverse effect, such as side effects, allergic reaction and the early development of infection [11, 13].
Discover new treatment opportunities: Utilize the enormous amount of data on past patient treatments and outcomes to improve patient care and provide personalized treatment.
One can easily access the real-time information on patient conditions and medical knowledge to make the right medical decisions as fast as necessary.
Leverage rapid analysis and simulations of clinical data, outcomes, and cost data to use your resources in a focused way to achieve optimal results.
Big data is extensively accepted to convert medical claims payment systems fundamentally. It is resulting in reduced submissions of improper, flawed or fake claims [11].
Challenges of Big Data in Healthcare [14]:

1. Failure to accept predefined data standards to declare interoperability across others settings.
2. Inferring knowledge from complex heterogeneous patient sources. Leveraging the patient/data correlations in longitudinal records.
3. Lack of willingness to engage in information sharing.
4. Understanding unstructured clinical notes in the right context.
5. Efficiently handling large volumes of medical imaging data and extracting potentially useful information and biomarkers.
6. Concerns about privacy and security.

III. SENTIMENT ANALYSIS IN HEALTHCARE: A BIG DATA APPROACH

Big Data in healthcare is associated with the exploding volume of patient-specific data. Patients are the foundation to take charge of their personal health care by performing research about their injuries and illnesses. They can also join social networks on which they can exchange information and provide support to each other. Big Data is also serving consumers more reliable and timely information about the quality of health care. The primary objective of Big Data Analytics in medicine is to improve the efficiency and quality of health care service. Patients can be treated successfully by analyzing diseases at earlier stages. It also helps in managing specific health populations and individuals and detecting health care fraud more quickly and efficiently.

Skuza, M. et. al. design a model which predict the price of stocks based on social media services. Machine learning was implemented for classification of sentiments [15]. Tedeschi, A. et. al. proposed a cloud based sentiment analysis framework for brand monitoring [16]. Georgiou, D. et. al. made a comparison between non commercial and commercial tools suitable for health care data[17].

Sentiment analysis in healthcare is enormously trending. Its assistance in healthcare organizations can improve the patient experience. By amalgamating patient data with their satisfaction value and the sentiments one can provide deeper insights in health care domain. This analysis also gives an provides the drastic change and their impact on the patient perception and values. Hybrid polarity detection is best usages for the sentiment summarization [18].

Sentiment analysis applied to patient data is a systematic study of online tweets and patients' reviews on satisfaction surveys. Tweets and patients' reviews are broken into small bags of words. These bags of words are further analyzed and classified according to their topic, meaning and intensity-whether and how positive or negative.

IV. METHODOLOGY

This section starts with a brief outline of the Twitter micro blogging system. Afterward, it presents the dataset used for assessing the performance of the proposed system. Then workflow is briefly explained. Results are then submitted and discussed.

A. A Briefing on Twitter

Twitter is a trendy micro blogging social networking service that is used to express feelings by reading and writing short messages on any topic. Tweets with a maximum extent of 140 characters are used more informally, with slangs and special characters. The accessing and monitoring of posted messages are possible through the Application Programming Interface (API) of Twitter. This API provides methods for data extraction that is tweets across the Twitter account. Internally, the Twitter API is divided into Stream API and Search API. The Stream API allows access for the flow of real-time tweets. The Search API provides access to a limited set of recent tweets. [19].

Here, we are dealing with Search API for extracting tweets containing particular search keywords. Twitter implements OAuth 1.0A as its standard authentication mechanism, and to use it to make requests to twitter's API, go to https://dev.twitter.com/apps and create a sample application as shown in fig.1.

There are four primary identifiers for an OAuth 1.0. These are consumer key, consumer secret, access token, and access token secret. For this any ordinary Twitter account is used to login, creation of an app, and for credentials.

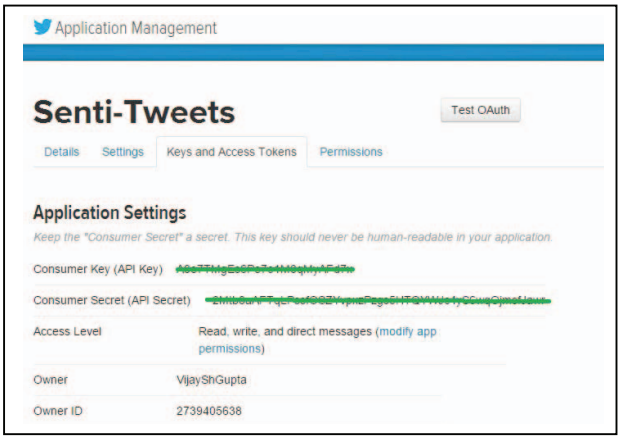

Fig. 1. Creation of Twitter Application

B. Dataset

The initially pilot study was conducted through an online questionnaire survey to get the patient data based on the diseases they were suffering. The hospital they preferred during illness, location, the age group they belonged to and the rate of satisfaction. Glimpse of data collected is shown in fig.2.

First Name	Age	Disease for which you visited the hospital	Name of Hospital	Satisfaction Rate	Place of Hospital
Vijay	26-35	Brain and Nervous System Diseases	Max Hospital	4	Delhi NCR
Purti	16-25	Injury	AIIMS	5	Delhi NCR
Saloni	26-35	Injury	AIIMS	5	Delhi NCR
anshul jain	16-25	Heart, Lung and Other Organ Diseases (ENT)	medanta	5	Delhi NCR
anshika jain	16-25	Other	Apollo	2	Delhi NCR
Damanjeet	26-35	Pregnancy and Childbirth-Related Diseases	Fortis	5	Delhi NCR

Fig. 2. Sample of Data gathered from survey

API helps to choose the desired language to retrieve tweets. English.is used as a default language as it is widely used by the developer group. Thus, the classification of the tweets work only for the tweet which are written in English language because the training data is in English only.

The data used for this paper contains 25000 tweets, 5000 tweets of every hospital. For positive sentiments, each row is

marked by 1 and 0 for negative sentiment. Details about sentiment is discussed later in this paper

Along with the twitter data, the system also requires other datasets like stop words, a dictionary of negative and positive words, an emoticon dictionary and an acronym dictionary for Twitter slang words.

C. Workflow and Packages used

For extracting tweets from Twitter, Twitter API and Hadoop framework is used. Hadoop Flume is used to capture tweets from the twitter and stores directly into Hadoop Distributed File System (HDFS). Elements of the workflow are:

1. Twitter API: Data can be collected from various sources like online and offline surveys, questionnaires, feed-backs, social networking sites, blogs, customer reviews, etc. based on some keywords. Authorization is required for this process if social networking sites like Twitter, Facebook, and LinkedIn, etc. are used. Different APIs are available which can be implemented easily while data collection. Twitter API is used for collecting tweets. For this a Twitter application is created shown in fig.1. After creating an application, a unique consumer key, consumer secret, access token and access token secret is generated. These keys are further used for the authorization of application so required tweets are extracted easily.

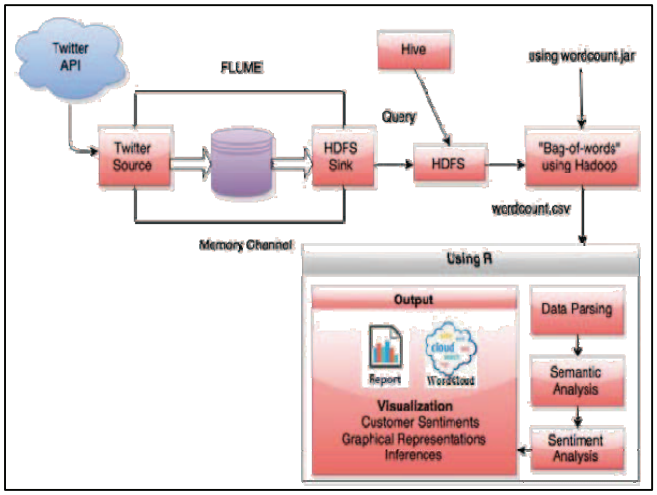

Fig. 3. Workflow of the System

2. Configure Flume: Now the flume-sources-1.0.jar is used. It contains the java classes which pull the Tweets from the twitter account. The pulled Tweets are stored in HDFS.
3. Configure Agents: The consumer key, consumer secret, access token and access token secret obtained are used. Here, name of the hospitals are used as Keyword for the extraction of tweets. The hdfs.path must point to the NameNode and the location in HDFS where the tweets will be saved.
4. Start flume entering the next command: The default NameNode Web Interface is available at http://localhost:50070/. And also tweets are placed in user/flume/tweets folder.
5. Configure Hive: The conf/hive-site.xml is modified with the locations of the NameNode and the JobTracker. Twitter tweets are in JSON format, so hive-serdes-1.0.jar is used to help Hive in understanding JSON format.
6. Create the tweets table in Hive: Tables can be created in Hive as:
   CREATE EXTERNAL TABLE tweets (id BIGINT, source STRING, user STRUCT < name: STRING, friends-count: INT, verified: BOOLEAN>) ROWFORMAT SERDE "com.cloudera.hive.viz:jasonasd" LOCATION '/usr/flume/tweets'
7. Playing with Hive: Now, the data is in HDFS and table is created in Hive so various queries can be run on the data such as:
   Hive > select user.name, user.friends-count c from tweets order by c.
8. Bag of Words using Hadoop: The numbers of occurrences of each word are counted using wordcount.jar in Hadoop.
9. ggplot2: This package provides widely used scenarios for the implementation of graphics in R. Lattice and base graphics provides the helpful ways to succeed the graphical hurdles.
10. RColorBrewer: This package provides a colourful way to the visualizations. A range of palettes are there to make vivid graphs and charts as per the variables.
11. tm package: This provides a text mining applications within R.
12. Wordcloud: This package helps in creating interesting and attractive wordclouds in R using text mining.

Step by step flow diagram of the proposed system is shown in fig.3. The packages discussed are used for the implementation of the proposed system. After data acquisition (tweets) through twitter API, data preparation and conversion of tweets into key & value pairs are done. Here Key is the word and values are the counts (occurrence of each word). These key & values are saved into the wordcount.csv file.

Once the wordcount file is generated, it is passed into R and several texts cleansing process is applied. After cleaning semantic and sentiment analysis is performed with the help of positive and negative dictionary file. Results and inferences are created in a form of wordclouds and bar graphs.

D. Evaluation Metrics

Once we have the collected tweets using Twitter and Hadoop API, applying some analytics to these tweets to visualize some kind of useful information. The ultimate motive of sentiment

analysis classify the words into two main category. The categories are positive and negative. For classification of the negative and positive words we need a list of words that all possible positive and negative sentiment words.

The overall sentiment is calculated by using sentiment function which first calculates the positive score and the negative scores by comparing words with the positive and negative list of words. The bar charts of these sentiments score are shown in fig.4, fig.5, and fig.6. The final score is calculated as:

Score= positive score - negative score

If Score > 0, this shows the Positive Opinion about the sentence.

If Score < 0, this shows the Negative Opinion about the sentence.

If Score = 0, this shows the Neutral Opinion about the sentence.

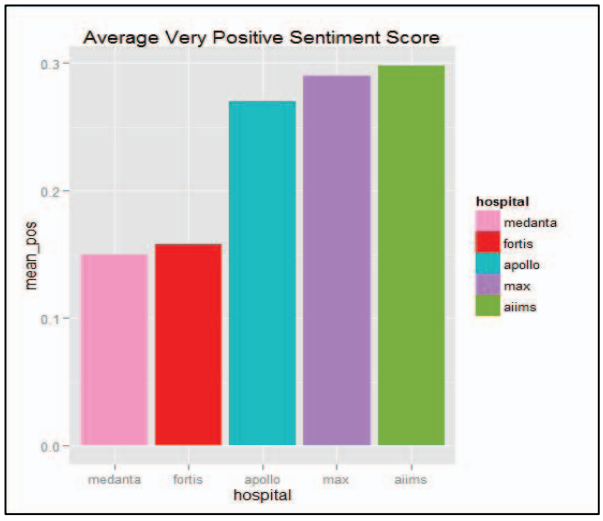

Fig. 4. Average Very Positive Sentiment Score System

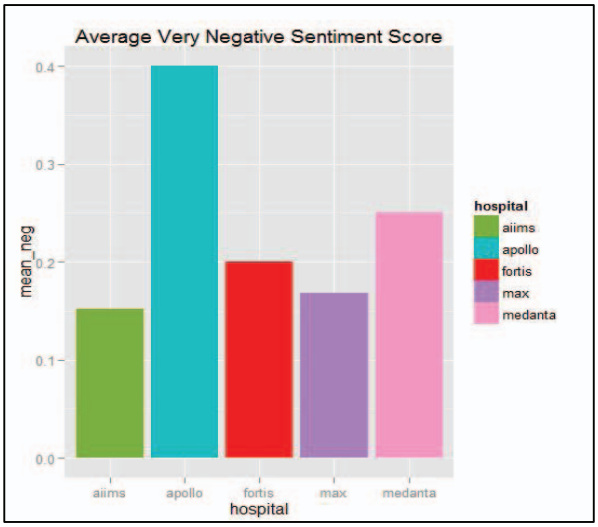

Fig. 5. Average Very Negative Sentiment Score

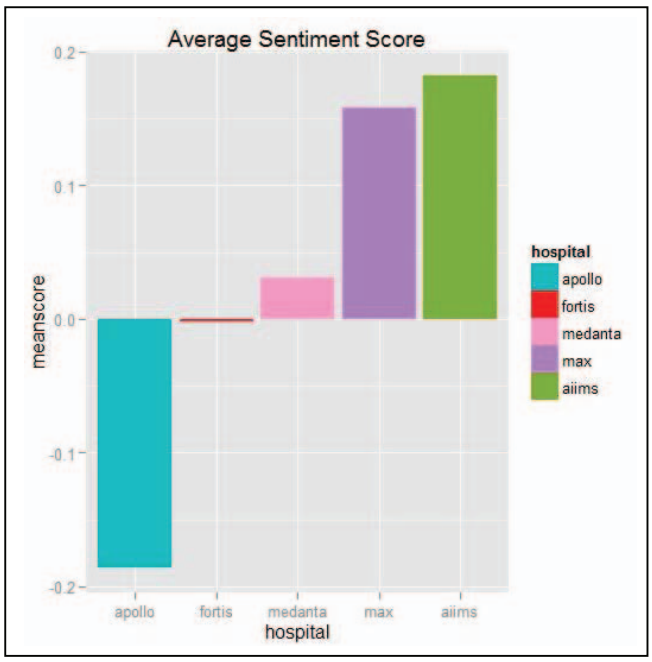

Fig. 6. Average Sentiment Comparison

E. Experimental Results

Initially, a pilot study was conducted through an online questionnaire to find out the popularity of different hospitals among patients. A sample of the data collected in the survey is shown in fig.2.

This study helped in selecting the hospitals on the basis of the satisfaction rate. The five most popular hospitals were chosen from the list that included: AIIMS, Apollo, Fortis, Max, and Medanta.

Tweets are collected for all the hospitals from Twitter using Twitter API. Once the tweets are gathered then they are cleaned and their semantics are analyzed.

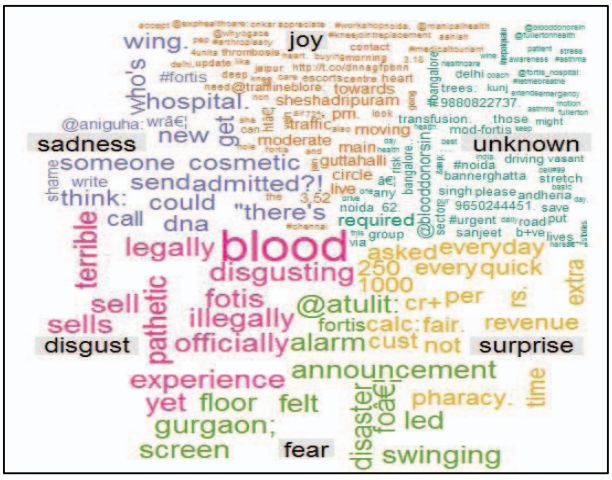

Fig. 7. WordCloud of Tweets for Fortis hospital based on emotions

Sentiment analysis for all the hospitals is conducted on the basis of emotion and polarity. Then a wordcloud is created on the basis of emotions recognized for the hospitals. Wordcloud of Fortis and Max hospitals are shown in fig.7and fig.8.

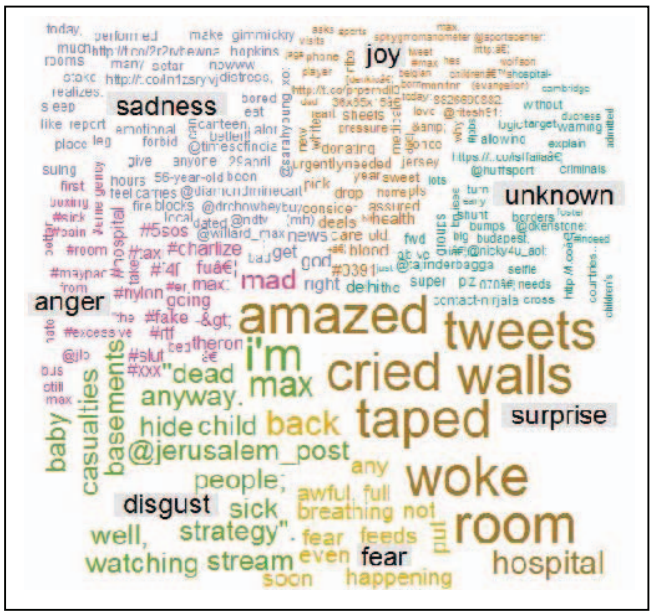

Fig. 8. WordCloud of Tweets for Max hospital based on emotions

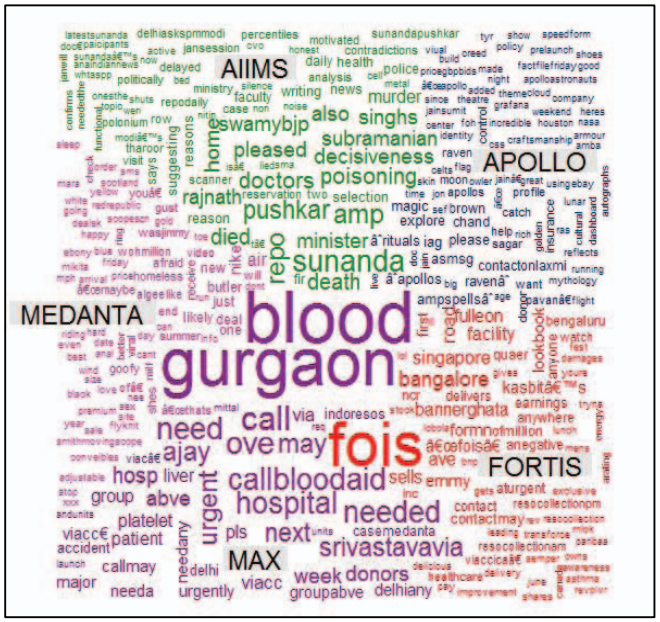

Fig. 9. Comparison WordCloud for five selected hospitals emotions

Once the score of each According to the comparison wordcloud shown in fig.9, we analyzed those different hospitals that provide better facilities and their level of popularity among patients at the time of illness.

Comparison wordcloud shows the efficient way to visualize the text context based on the frequency of word and their sentimental values. In fig.7 & 8, sentiment wordcloud is used to demonstrate the different sentiment emerges from the tweets of the people for Fortis and Max hospitals. It shows the voice of public in the form of wordcloud. The frequency of the occurrence of particular sentimental words can be explored using visual bars, box plots, histograms, scatter plots. This can be done by using different functions available in R. Here ggplot2 package is used which contains geombar as a function of creating bars.

V. DISCUSSION

Big data analytics shows the potential to reduce the complexity of the health care domain. As per the online survey, it was observed that the patients have trust among some hospitals like AIIMS, Apollo, Max, Medanta, and Fortis. People prefer these hospitals for the quality of service provided and the facilities offered to them.

Then on the basis of the satisfaction rate the data related to these hospitals is collected from Twitter. The so collected data is used to plot graphs and for creating Wordcloud.

By the comparative study performed it was observed that the green bar of AIIMS had maximum positive sentiments among all the hospitals. Thus, it is the most encouraged hospital while the blue bar of Apollo hospital has maximum negative sentiments, thus it is the most discouraged hospital. The second best hospital according to the results is Max.

Thus, the ranking of the hospitals based on sentiment analysis is AIIMS, Max, Medanta, Fortis and Apollo in the ascending order. There are some limitations regarding the Search API used for data extraction that we should address, which are:

1. A fixed number of tweets are extracted using twitter API at a time.
2. While extraction of tweets it may possible that the number of tweets fetched is less than requested.
3. Older tweets are not fetched with this extraction procedure.

VI. CONCLUSION AND FUTURE SCOPE

People are stated using new application and tools for the better health care. They also validate the doctor prescriptions and essential tests required for the particular disease. The medical domain becomes more effective and efficient. Big data creates a revolution to transform the look of healthcare domain and significantly improve clinical outcomes.

Perhaps, the greatest concern among the public is regarding health and therefore it has become necessary to analyze the opinions of the users posted on the networking sites so that better health care can be provided.

There is a significant influence of social media on the life of people. The positive, negative influence can be used to depict real-world phenomena or can be associated with real-world events. Use of social media is tremendously increased among

users, through the surveys, tweets, comments, blogs, and feed-backs. We depicted the correlation between the popular hospitals and their patients and through the sentiments generated we can create a framework for the hospital recommendation system.

In this work we evaluate the sentiments of 5 major hospitals of the Delhi NCR region. These hospitals were selected based on the survey results. After that their tweets are analyzed and visualized. On the basis of the sentimental analysis of tweets, we concluded that people of Delhi NCR region have maximum positive thought or sentiments for the AIIMS and least for the Apollo. This is also being visualizing better in fig.6.

This work demonstrates that the comments and view of patients are a novel approaches for the development of the hospital recommendation system. A traditional method that is a survey helps in concluding the name of the hospital for the further research. Although Detection of sarcasm, as well as, exploring more features would be additional future works of this study. The development of hospital recommendation system is a future direction of this paper.

REFERENCES

[1] Ji, Xiang, Soon Ae Chun, and James Geller. "Monitoring Public Health Concerns Using Twitter Sentiment Classifications." In Healthcare Informatics (ICHI), 2013 IEEE International Conference on, pp. 335-344. IEEE, 2013

[2] Bollier, David, and Charles M. Firestone. The promise and peril of big data. Washington, DC, USA: Aspen Institute, Communications and Society Program, 2010.

[3] Bill Hamilton, Big Data is the Future of Healthcare, Cognizant 20-20 Insights, September, 2012

[4] Michael Wessler, OCP and CISSP, Big Data Analytics for Dummies (Alteryx Special Edition), Wiley Publisher

[5] Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela H. Byers. "Big data: The next frontier for innovation, competition, and productivity." (2011).

[6] Mathiesen, J., Angheluta, L., Jensen, M. H., "Statistics of cooccurring keywords in confined text messages on Twitter", The European Physical Journal Special Topics, vol. 223, issue 9, September 2014, pp. 1849-1858

[7] Gantz, John, and David Reinsel. "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east." IDC iView: IDC Analyze the Future 2007 (2012): 1-16.

[8] Networked European Software and Services Initiative (NESSI) Big Data White Paper, Big Data A New World of Opportunities, December, 2012.

[9] Jain, Purti, and Shruti Kohli. "Big Data Analysis, Algorithms and Applications: A Survey", International Journal of Emerging Trends in Engineering Research ,Vol. 3, 2015, pp. 19 – 24

[10] Kogent Learning Solutions, Bill Franks, Wrox Certified Big Data Analyst (WCBDA), Introducing Big Data Analytics and Predictive Modeling, Wiley Publishers.

[11] Cottle, M., Hoover, W., Kanwal, S., Kohn, M.,Strome, T. and Treister, N.W., Transforming Health Care Through Big Data, Institute for Health Technology Transformation(IHTT), 2013.

[12] Raghupathi, Wullianallur, and Viju Raghupathi. "Big data analytics in healthcare: promise and potential." Health Information Science and Systems 2, no. 1 (2014): 3.

[13] Data, Demystifying Big. "A Practical Guide to Transforming the Business of Government." TechAmerica Foundations Federal Big Data Commission (2012).

[14] Sun, Jimeng, and Chandan K. Reddy. "Big data analytics for healthcare." In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1525-1525. ACM, 2013.

[15] Skuza, Michal, and Andrzej Romanowski. "Sentiment analysis of Twitter data within big data distributed environment for stock prediction." Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on. IEEE, 2015

[16] Tedeschi, A., and F. Benedetto. "A cloud-based big data sentiment analysis application for enterprises' brand monitoring in social media streams."Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI), 2015 IEEE 1st International Forum on. IEEE, 2015.

[17] Georgiou, Despo, Andrew MacFarlane, and Tony Russell-Rose. "Extracting sentiment from healthcare survey data: An evaluation of sentiment analysis tools." *Science and Information Conference (SAI), 2015*. IEEE, 2015

[18] Bahrainian, Seyed-Ali, and Andreas Dengel. "Sentiment analysis and summarization of twitter data." In Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on, pp. 227-234. IEEE, 2013.

[19] Lima, Ana CES, and Leandro N. de Castro. "Automatic sentiment analysis of Twitter messages." In Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on, pp. 52-57. IEEE, 2012.