Predicting Elections Using Twitter Data Analysis

Saurabh Roy, Saurabh Pandey, Himanshu Khairajani

***Abstract* — With the rapid development of the Internet, more and more users are expressing their views using social media platforms and hence social media have changed the information behavior because of Realtime access of information to the people without restriction of time and data resources. About 1 quintillion bytes of data is generated on the daily basis, enriched of sentiments in the form of posts, status updates, tweets and blog posts etc. In this project, we are going to analyze a large number of election-oriented post by general social media (twitter) user. Based on these users posts, system will analyze if there exist any pattern between these posts and will apply sentiment classifier model and draw relevant information from the collections of these posts recorded over some period of time; the presented method point out the likelihood of development of a categorization model to identify the political orientation of the social media users based on the posted content and other social media based traits. Now since the twitter dataset which is needed to be extracted and processed for sentimental classification followed by the election result prediction contains over billions of tweet texts from twitter users, hence is beyond the capabilities of the normal processors. That is where Big Data concepts and methodologies comes into the picture. We will use popular Big Data tools like Apache YARN, Apache HIVE, Apache HADOOP (HDFS and MapReduce), Apache Flume etc. in order to efficiently extract, manipulate and process such massive amount of data and will use this analysis for the election prediction.**

***Index Terms*—Big Data, HADOOP, HIVE, Flume, YARN, HDFS, MapReduce.**

I. INTRODUCTION

TODAY'S era is an era of Internet. Every answer can be found here. The people are now connected through it. While talking about the internet, social media platforms play an important role by connecting peoples all across the globe. Youth spends a lot of time on platforms like Instagram, twitter, YouTube, short video app, Facebook etc.

Social media platform is a mobile app and/or web-based system that provides the creation, access and exchange of user generated information. Social platforms are very crucial for researchers in computational social science, investigating the queries using quantitative methods like ML, computational statistics and enamored data handling technologies like big data for simulation modelling and data mining. People analysts and businesses feel the need to gain new insights from social media; they need the tools of analytics and technology to revolutionize this huge amount of data. The Apache Hadoop software library is a framework [19] that allows for distribution of the operation of big datasets covering computer groups using elementary programming models [20].

As per the statistics, INDIA is the second largest user of Internet in the world after CHINA. This is due to tremendous revolution in telecom sector. The use of social media becomes riskier due to security, privacy, and harassment, but still it gives various opportunities to its users for knowledge sharing engagement and collaboration.

The platforms like Facebook, Instagram, WhatsApp, YouTube, twitter motivates mankind to turn engaged in political activity by sharing their opinion about candidates, parties and their previous work. Social media has a great impact on political elections. Social Media Analysis is a useful tool for obtaining details of the sentiment of customers distributed across online sources. Twitter, one of the biggest social media sites that receives millions of daily tweets of sorts important matters. The authors of those messages write about their lives, sharing ideas in different ways, topics and discussing possible problems. This post analysis can be used for internal decision-making in various areas like government, Election, Business, and Product reviews etc. Doing Sentiment Analysis on Twitter is more complicated because tweets are too short (only 140lcharacters) and they usually contain slang, Icons, hashtags and other twitter-specific twitter sticks. For the purpose of developing twitter provides a streaming that allows the developer to access 1% of the tweet sent at that time based on a specific keyword. Twitter data is usually a random use i.e. very brief summaries high. Also allows the use of icons that are direct references to the author's view on the title. What we want to do is emotion analysis on data being sent to twitter API's which makes mining progress and provides tweets only related to that.

II. LITERATURE SURVEY

The author Min Song, Minchul Kim and Yoo Kyung Jeong investigated a dataset collected from instantaneous tweets during election of Korea in 2012. Topics extracted from tweets and related instantaneous event relations were associated and were traced chronologically using co-occurrence retrieval

This journal is submitted for review on July 8, 2020. This work was supported in part by International Institute of Information Technology Naya Raipur.

Saurabh Roy is with department of computer science, International Institute of Information Technology Naya Raipur, Raipur, CG, 493661 India (e-mail: saurabhr17100@iiitnr.edu.in).

Saurabh Pandey is with department of computer science, International Institute of Information Technology Naya Raipur, Raipur, CG, 493661 India (e-mail: saurabh17100@iiitnr.edu.in).

Himanshu Khairajani is with department of computer science, International Institute of Information Technology Naya Raipur, Raipur, CG, 493661 India (e-mail: himanshu17100@iiitnr.edu.in).

techniques.

India for the 2016 general election; End user trend towards parties and candidates' study by author Abhishek Bhola [3] using Twitter. A dataset containing 22.85 million posts analyzed to identify user, candidate and party popularity by time, subject or location. There was a sentimental analysis performed using categorization algorithms.

In nation like Greece Voting Advice Applications (VAAs) is designed for election analysis. It is befitting more and more favorable and they are serving people to decide which party/Candidate to vote. It's built on region-based recommendation system. It anticipates a correlation of political opinion of users, and becomes a medium for end user conveying. This system gives various approaches proposed for region-based vote recommendation. The approach was estimated on five real VAA data sets in terms of prediction accuracy.

Utilizing data from Instagram, Facebookland Twitter; from Larsel Kaczmirek and his team GESIS collected many slant of communication system. They collate data collected from social platform with native surveys and compute new insights by providing social media, how it can be utilized in the course of elections. Based on this study, German Longitudinal Election Study (GLES), a deep-rooted research project is designed, which evaluate the 2009 German elections followed by predictions of 2013, and 2017 elections. The chief objective of this project is to trace the process of German elections across an amplified span of time by gathering Instagram posts, Facebook data and Tweets about German Bundestag election.

There are several tasks related to sentiment analysis. Unnamalia K, Ohbyung Kwon and Namyeon Leea [6] showed that the analysis of sentiment tweet are a challenging task due to multilingual and informal messages. In this work, the research model is presented to clarify the aim to acquire big-data analytics mostly from the conceptual view point of usage and data quality management experience. Our such empirical investigations show that a firm's intention for big data Analytics can be firmly affected by its potential to maintain quality of the corporate data.

Chetashri Bhadanea also with William D. Abilhoa and Leandro N. de Castro [7]. This page shows how to extract keywords from tweets collection that represent text as graphs and work with steps to get the right size vertices(keywords). To evaluate the effectiveness of the presented method, three contrasting test sets were used. The first test bids to TKG. A text from a time magazine and compares its effectiveness with that of a book. The second set of research takes posts from three different TV shows, it works TKG also equate it to TFIDF and KEA, which have demographic categories as benches. Finally, these three algorithms are used in posts of growth sets, its size and duration of competition are measured and compared. Overall, these tests yield a common rundown of how TKG can be utilized in practice, its own performance equated to other standard methods, and how it balances the larger data instances. The results indicate that TKG is a novel annotation and a strength remove keywords from text, especially in short messages, such as tweets [3].

Bo Pang and Lillian focused on requests for redevelopment in the best way possible, emits more sensitivity than one text. The scope of their application ranges from harmony of understanding in understanding text topics, visualization and the feeling and automation of customer review help [5].

Conceptual analysis has the most popular technology in the modern world. A lot of work has it done in this field. The following are the most popular methods in the modern world in this regard technology. There is a plenty of analysis in the area of analysis. BolPang and Lee were the pioneer king in this field. What is currently being done in this area includes use of a mathematical method that uses a formula for inflation depending on proximity of words with adjectives such as 'very good', 'bad', 'bad 'etc. The project uses the 1-Naïve Bayes method and the HADOOP cluster for distribution of data.

Performance of all data types.

Comparative studies among different methods have been reviewed, processed that include subjection detection, feature selection, and different machine learning methods [1]. Various machines have been used so far, which includes handbags, a training corpus, a literary level, a sentence level and level of mining level [3]. Various forms of prominence are becoming available in an external system where spiritual processing is used. Language feature and appropriate domain features are essential for providing better levels of

text [2]. So, in this program, consider the gamut of associated keywords

a trait is essential for successful isolation. The algorithm explained revolves around the expansion and better understanding of the model proposed by Dave, Lawrence and Pennock [8].

Turney et.al. [9] A bag-of-words approach in which word relations were never taken for emotional analysis and the sentence is simply it is considered a collection of words. Finding all the emotions in a sentence, the feelings of each word were determined differently values are computed using other integrating functions.

Pak Paroubek [10] suggested the model of classifying posts as good and bad. By utilizing the Twitter API, they formed a twitter corpus by gathering tweets and clarifying that Tweets use thumbnails. He uses that company and a diversified view of the Naive Bayes a classifier method was developed using features such as POS-tags and N-gram. The training set used in the test does not work well because they only collected Tweets with thumbnails.

Po-Wei Liang et.al. [11] [12] used the Twitter API in collecting data from twitter. Tweets contain filtered views. The Unigram Naive Bayes model was developed to identify polarity. They have worked to remove unwanted signals by using Mutual Information and Chi, the method of the square subtraction factor. Ultimately, how to predict tweets as Positive or negative did not provide better accuracy in this way. That [4], suggests a

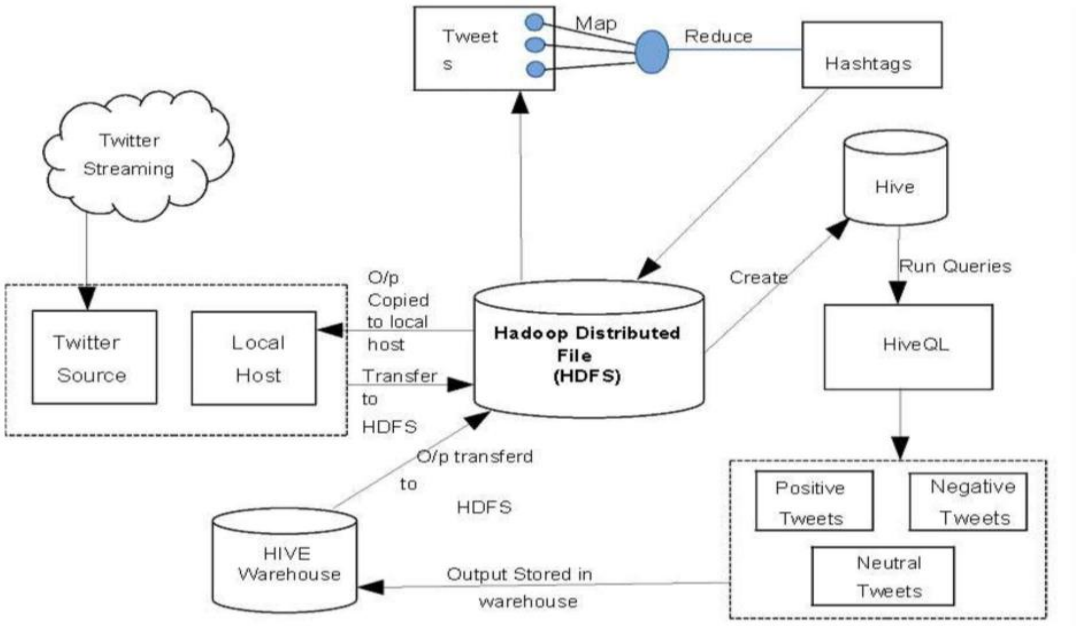

Fig. 1. System Architecture.

a language-based mining assistance program, which clause the Stage analysis of comprehension of the analyzed texts. For every message a post sentence produces a valid dependency tree, and then includes a sentence in phrases. It then determines the context of the understanding based on the location of each paragraph. The grammatical dependency of words also uses Sentiments WordNet which had previous feelings Word scores also come from domain specific dictionaries.

Hussein [13]; the paper describes past has the purpose to associate the various important problems in how we feel and explore how to raise the accuracy outcome suitable for the techniques used.

All of the above work uses Corpus data. But in this paper, we use the actual broadcast data which were preprocessed based on the text filters used and doesn't require storage for storing tweets.

I. METHODOLOGY AND APPROACH

Our model is based on the technique that analyses the posts on the basis of keywords and hashtags. The implemented system collected the data on the basis of hashtags related to political parties and their agendas. The political orientation of peoples toward any particular party/ candidates can be predicted using these posts on social media. Now a day's social media is flooded by journalists, politicians, film stars and academicians; for its extremely political value. Many politicians and political parties use these platforms to spread their agendas to general peoples specially the youth. These all post can be categorized on the basis of various points like to get the people opinion of any particular geo-location / area which might be helpful for the parties to design their political campaign for victory. This proposed model basically focused on collecting the posts to use volume analysis. A trend analysis on a popular and rending political parties/ candidates and a sentiment analysis to distinguishing on negative positive and posts for a candidates and party help them to act accordingly to make their reputation high. These trends also help the people to make their opinion about the political parties and their candidates. Our system architecture for this work is depicted in fig. 1. The detailed description of the steps for the analysis is given below.

A. Dataset

For sentiment analysis, we have combined together two datasets

The first one is the IMDB movie review dataset as a single customer review training dataset which is publicly available at Keras. This dataset includes 25,000 movies reviews from IMDB, labelled by sentiment (positive/negative).

Election related tweets from twitter database will be the second dataset for the system. Thesaurus will be given for Sentiment analysis (positive/negative)

. The dataset looks like Fig. 2.

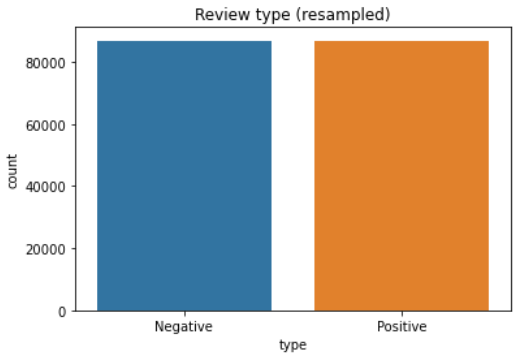

Fig. 2. Distribution of reviews among different classes.

B. Algorithms Used

Following is the algorithm for analysis:

Algo 1: Porter Stemmer Algorithm

Input

·Let G be downloaded tweets' set

Output

· Organized tweets with every unnecessary special character, words and space being removed.

Algo 2: K-Means Clustering Algorithm

At first, subject sensible tweets are put as a lope of clusters. The separation between sample and center is checked and pattern is appended to respective cluster in each iteration. Between sample and center, distance is calculated using TF-IDF and depending upon its weightage the cluster are updated at every iteration.

Input

· Let C be Set of centers where C = c1, c2, …. cc

·Let X be Set of data points where X = x1, x2, x3, … an

Output

· Established clusters

Steps

1: No. of clusters to be determined are chosen first

2: Randomly select the centroid for initial centers of the clusters.

3: Repeat

3.1: Using Lowenstein distance assigned to every object to their closest cluster center.

3.2: Calculating the mean points to compute new cluster center.

4: Until

4.1: object's clusters aren't changed further

OR

4.2: No more changes in the cluster's center

Algo 3: Naive bays Classifier

Topics can be analyzed from tweets regarding political orientation of users towards any party. To classify tweets into positive, negative and neutral classes a Map-Reduce version of naïve Bayes algorithm will be implemented.

Steps

1: Create data for the classifier

1.1: Construct an array consisting of positive tweets

1.2: Construct an array consisting of negative tweets

1.3: Combine these two arrays in a single list with two parts, for each tweet and its respective type.

2 Design Classifier

2.1: From the list extract the word feature list along with the count of its frequency.

2.2: Using this extracted list construct a feature extractor containing the words. These words are then compared with dictionary (which we created) indicating the passing or failing of the same.

3: Then training dataset is used to train the Classifier.

3.1: Generate Lable_Prod List containing negative and positive labels.

3.2: Generate Feature_Prod List containing the featured words.

4: Positive and negative Label probability is then Calculated.

5: Finally, to know the category of the tweet as negative, positive or neutral, comparer this probability.

C. Hadoop HDFS

The Hadoop Distributed File System (HDFS) is the primary data storage system used by Hadoop applications. It uses Name Node and Data Node to implement a distributed file system that is use to provides faster access of data across highly scalable Hadoop clusters. Hadoop is based on the master/slave architecture. Master node manage the metadata file while slave node stores the actual data. Hadoop is most reliable and efficient for processing of large amount of data.

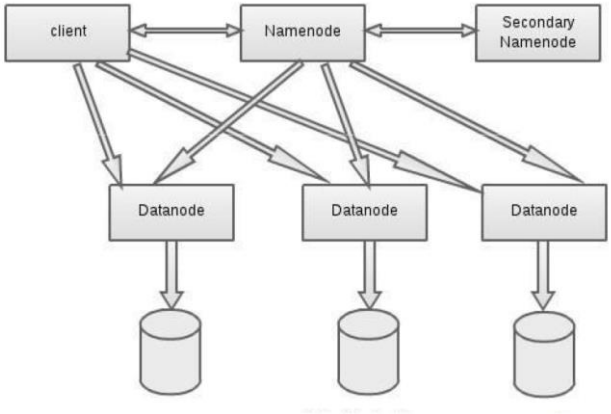

Fig 3: HDFS

D. Hadoop MapReduce

Map Reduce is a parallel processing software use to run on large cluster. MapReduce works in the coordination of two function map () and reduce (). Map split the data into tuples i.e. (key/value pair) sort them then feed this data as an input to the reduce () function. The input and output of the job is stored in HDFS. Scheduling of task, execution failure of job and monitoring is carried by this framework. Map reduce is mainly design to process large sum of data efficiently and reliably.

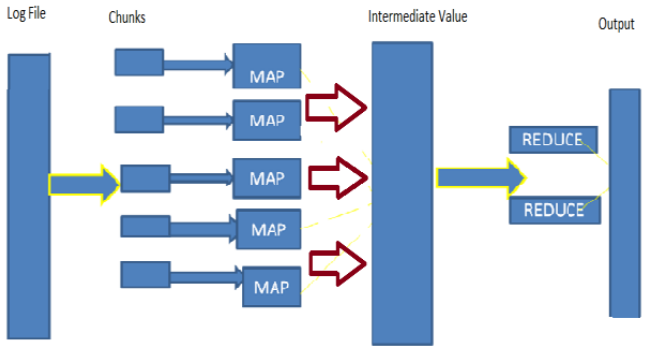

Fig. 4. Structure of Hadoop MapReduce

E. Apache YARN

YARN ("Yet Another Resource Negotiator") works as a resource manager for the data cluster in Hadoop. YARN is equipped with different data-processing techniques viz batch processing, stream handling, graph computations, interactive threading, and so on, in order to retrieve, run and process all the data clustered and stored in Hadoop Distributed File System [HDFS]. Also, YARN acts as a central platform that provides job scheduling, data governance, security which is not only limited to the MapReduce but also for numerous other application such as Spark, Hive, HBase, etc. YARN can also be integrated with various popular programming languages such as C++, Python and Java. Therefore, YARN opens up Hadoop to other types of distributed applications beyond MapReduce. The major components of Apache Hadoop YARN Architecture are:

1. Resource Manager: YARNs resource manager executes on the master node and is responsible for allocation of resource in the Hadoop cluster.

2. Node Manager: YARNs node manager executes on the slave nodes and it is their duty to execute a task on each single Data Node.

3. Application Master: They manages all the jobs assigned via user and grants resources on demand of separate applications. They exert together with the YARNs Node Manager and look after the proper conduction of scheduled tasks.

4. Container: It contains a bundle (package) of single node resources that includes CPU, RAM, HDD, Network, etc.

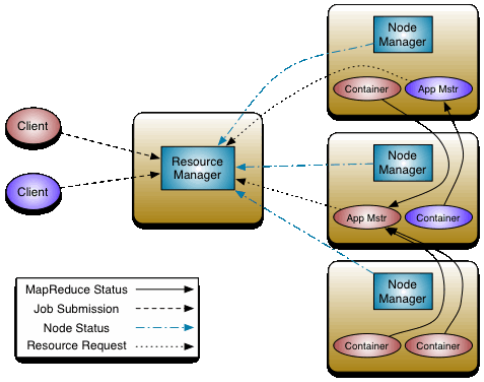

Fig. 5. YARN architecture.

F. Apache Flume

Apache Flume is a highly reliable and distributed tool which can be deployed in HDFS for data ingestion services. It facilitates genuine, portable, and distributed service which efficiently collects, aggregates, and transports a massive proportions of event/stream/log data from a numerous source like clouds, web servers, etc. to HDFS central data cluster or any other sinks like HBase. Apache Flume is composed of a simple and pliable architecture designed according to the streaming dataflow. It is equipped with tunable failure-recovery management system which makes Flume a vigorous and fault tolerant service. It employs a flexible and extensible data model that features online analytic application. In this project, we have used Apache Flume agent in order to collect the ongoing huge amount of streaming data related to elections from Twitter database through Twitter Agent event handler and store them to HDFS sink.

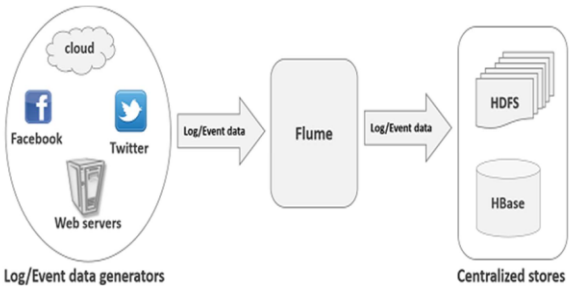

Fig. 6. Flume architecture.

G. Apache Hive

Apache Hive is a reliable data warehousing software which is used on the top of any Distributed System like HDFS hence making easy and effective querying and analyzing of structured data. It is familiar, fast, scalable and extensible. This robust tool is used for all ETL (Extract Transform Load) manipulations on database through HiveQL language which has interface similar to SQL due to which it is gradually becoming the top choice especially for Hadoop Big Data Analysis. In this project, we have used Hive for storing and handling the twitter data related to elections and then using the processed data for exit poll predictions.

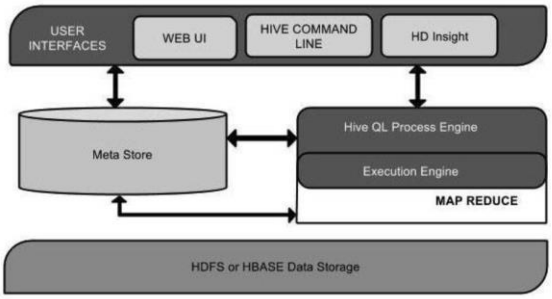

Fig. 7. Hive architecture.

II. Results

Based on Hadoop platform and java for Map Reduce framework, we have implemented this system. Porter Stemmer Algorithm and other user-defined functions are used to process the tweets from twitter database. And as discussed above these processed tweets will be pre-owned as input for the system for various analysis modules to generate sentiment, trend and volume analysis. We have applied our model on the above-mentioned twitter election data of Maharashtra 2017 general election and achieved an accuracy result which is consistent with the actual result election.

The complete analysis of our model result is as follow:

Hashtag's	Count	Hashtag's	Count
#NarendraModi	175	#APP	60
#namoagain	100	#RSS	45
#Rahulgandhi	139	#kejriwal	89
#BJP	126	#SoniaGandhi	100
#INC	68	#PriyankaVadra	95

Table 1. Hash Tag Wise Tweet

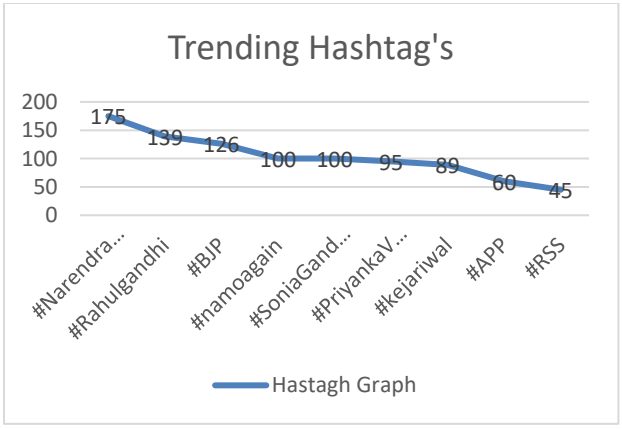

Fig.8 Hashtag Wise Tweet Count

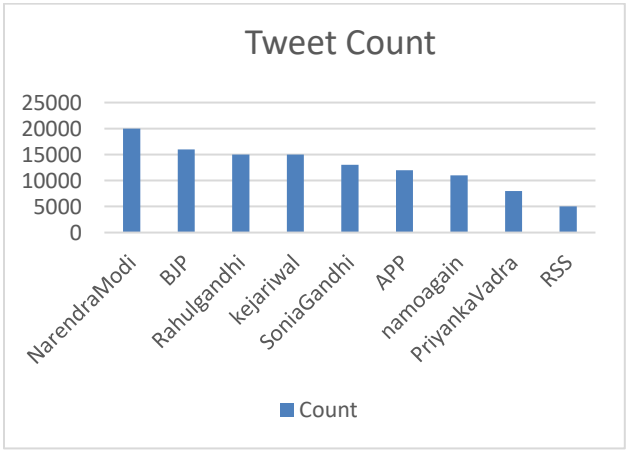

Fig.9 Trending Topic of Elections

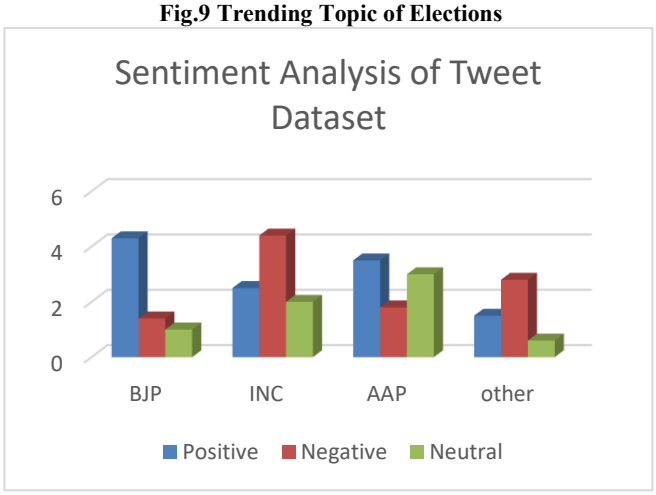

Fig.10 Sentiment analysis of tweet dataset

III. Conclusion

Big Data is really a fascinating concept which have an endless scope with infinite applications. Observing the expanded use of social media platforms, this cutting-edge paper concentrated on exploring of social platform as the chase for elections campaign. Understanding India to be one of the highest socially connected countries, having greater than 70% of its young generation below 35 years of age; Social platform plays essential part young youth's life. The designed system had worked upon the analyses of the Maharashtra state meeting election; taking a look at the impact of social platforms on the politics system of Maharashtra, found people can express their perspectives in one hundred forty characters more efficiently and openly. Our model was found to be accurate which is consistent with the actual result election. Hence, we can safely conclude that the Big Data tools like Hadoop HDFS, Hadoop MapReduce, Apache YARN, Apache Flume, Apache Hive, etc. are very efficient and reliable to analyze and deal with huge data applications with a good accuracy, which is impossible to achieve with traditional systems.

Acknowledgment

We would love to express our deep sentiments of gratitude to all who rendered their precious steerage for this work. We would like to thank Dr N. Srinivas Naik, Assistant Professor, Computer Science and Engineering Department, IIIT Naya Raipur, for guiding us with the topic and also providing us with adequate facilities, approaches and all the resources which were immensely needed to complete this paper.

References

[1] India is now world's second largest Internet user after China. https://www.statista.com/topics/2157/internet-usage-in-india/ .

[2] Gayatri Wani , Nilesh Alone, "A Survey on Impact of Social Media on Election System" http://www.ijcsit.com/docs/Volume%205/vol5issue06/ijcsit20140506100.pdf .

[3] Social media for political campaign in India. http://www.slideshare.net/RaviTondak/social-media-for-political-campaign .

[4] Min Song MeenChulKim ;Yoo Kyung Jeong, Analyzing the Political Landscape of 2012 Korean Presidential Election in Twitter 1541-1672/14/ Published by the IEEE Computer Society..

[5] AbhishekBhola "Twitter and Polls: Analyzing and estimating political orientation of Twitter users in India General Elections2014" arXiv:1406.5059 [cs.SI].

[6] IoannisKatakis, Nicolas Tsapatsoulis, Fernando Mendez, VasilikiTriga, and ConstantinosDjouvas "Social Voting Advice Applications - Denitions, Challenges,Datasets and Evaluation" IEEE TRANSACTION CYBERNETICS ,VOl. 44 No. 7.

[7] Use and Rise of Social media as Election Campaign medium in India,Narasimhamurthy N,(IJIMS), 2014, Vol 1, No.8, 202-209 .

[8] Indian general election, 2014. http://en.wikipedia.org/wiki/Indian_general_election,_2014 .

[9] Population of India 2015. http://www.indiaonlinepages.com/population/india-current-population.html .

[10] Twitter Data Analytics. http://tweettracker.fulton.asu.edu/tda/TwitterDataAnalytics.pdf .