Predicting Elections Using Twitter Data Analysis

Himanshu Khairajani (17100017)
B.Tech – VI Semester
Department of Computer Science and Engineering
DSPM International Institute of Information Technology - Naya Raipur
Atal Nagar, Chhattisgarh
Email: himanshu17100@iiitnr.edu.in

Saurabh Pandey (17100049)
B.Tech – VI Semester
Department of Computer Science and Engineering
DSPM International Institute of Information Technology - Naya Raipur
Atal Nagar, Chhattisgarh
Email: saurabh17100@iiitnr.edu.in

Saurabh Roy (17100050)
B.Tech – VI Semester
Department of Computer Science and Engineering
DSPM International Institute of Information Technology - Naya Raipur
Atal Nagar, Chhattisgarh
Email: saurabhr17100@iiitnr.edu.in

Under the supervision of

Dr. N. Srinivas Naik
Assistant Professor,
Department of Computer Science and Engineering
DSPM International Institute of Information Technology - Naya Raipur
Atal Nagar, Chhattisgarh
Email: srinu@iiitnr.edu.in

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.

1. 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 internet, social media platform plays and 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. All current topic is due to support of people views, comment, like and share that creates a clear view about that particular topic to the whole world. No doubt credit for all these goes to Social media platforms.

Social media platform is a mobile app and/or web-based system that provide the creation, access land exchange of user generated information. Social media is a very important for researcher in computational social science that investigate the questions using quantitative methods like machine learning, computational statistics and enamors data handling technologies like big data for simulation modelling and data mining.

ocial media has led to many data services, tools and analytics forums. The tools available to researchers are likely to provide more access to raw data. Researchers need to incorporate analytics into a language like Java. So, the proposed work ist better than the ones available in terms of cost, Big Data handling and scale stability. 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 they will have greater volume land variety in the appropriate techniques to reach certain conclusions.

The Apache Hadoop software library is a framework [19] that allows for distribution of the operation of large data sets across computer groups using simple programming models [20]. Provides high-risk and flexible properties with the same functionality. Instead of trusting on computer platforms to deliver high availability, the library itself is designed to find and address it is a failure of the application framework, thus delivering the most available service on top of the cluster computers, each of which can be prone to failure. Therefore, Social Media Analysis is a useful tool for obtaining details of the sentiment of customers distributed across online sources [18].

2. MOTIVATION

As per the statistics, INDIA is the second largest user of Internet in the world after CHINA. By survey of "Internet and Mobile Association of India" (IAMAI) in Feb 2019 INDIA have more than 480+ millions of active internet user which is 53% more as compared to year 2016.This is due to tremendous revolution in telecom sector. The use of social media become riskier due to security, privacy, and harassment, but still it gives various opportunity to its user for knowledge sharing engagement and collaboration.

The platforms like Facebook, Instagram, WhatsApp, YouTube, twitter motivates people to get engaged in political activity by sharing their opinion about candidate, parties and their previous work. Social media have a great impact on political election. The candidates plan their campaign strategies for the election from digital media feedback and voter get updates on political events from digital media.

3. AIM

Social analytics collects and analyzes consumer ideas and translate them into insights land help businesses identify areas customer satisfaction of any customer complaint. And it provides fast response to marketing campaigns, so that we analyze the campaign that will be well received by customers. Social analytics serves as a new channel between consumers land industries [21]. Also helping them provide a review of their product impact on the market. Hence the proposal the model meets the needs of companies bi analyzing data well and delivering results.

4. OBJECTIVE

Social Media Analysis is a useful tool for obtaining details of the sentiment of customers distributed across online source. Micro logging today has become a popular communication tool among Internet users. 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 discuss possible problems. This post analysis can be used for internal decision-making various areas like government, Election, Business, and Product reviews etc. Status analysis is one off the most important twitter gaming analysis sites that can be very helpful decision making. Doing Sentiment Analysis on Twitter is more complicated than doing it with major updates. This is because tweets are too short (only 140 characters) and they usually contain slangs, 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. What we want to do is emotion analysis is sent to twitter API's which makes mining progress and provides tweets only related to that. 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. Tweet messages also contain a timestamp land a username. This timestamp is helpful to guess the application of our future project.

5. RESOURCES USED

1.1 Social Media **developer account-**

- Developer account is need to be setup for Authentication key
- Once Authentication key is obtained use that key to get stream data

2.2 **Software requirements-**

- Hadoop framework is one the best framework for data analysis because of its flexibility to work on large scale distributed processing of data cluster.
-

3.3 Hardware requirements-

Hardware	Specification
CPU speed	2-2.5 GHz
Logical or virtual CPU cores	8-12
Total system memory	16GB
Local disk space for yarn.nodemanager.local-dirs	50GB
DFS block size	128MB
HDFS replication factor	3
Disk capacity	32GB
Total number of disks for HDFS	2

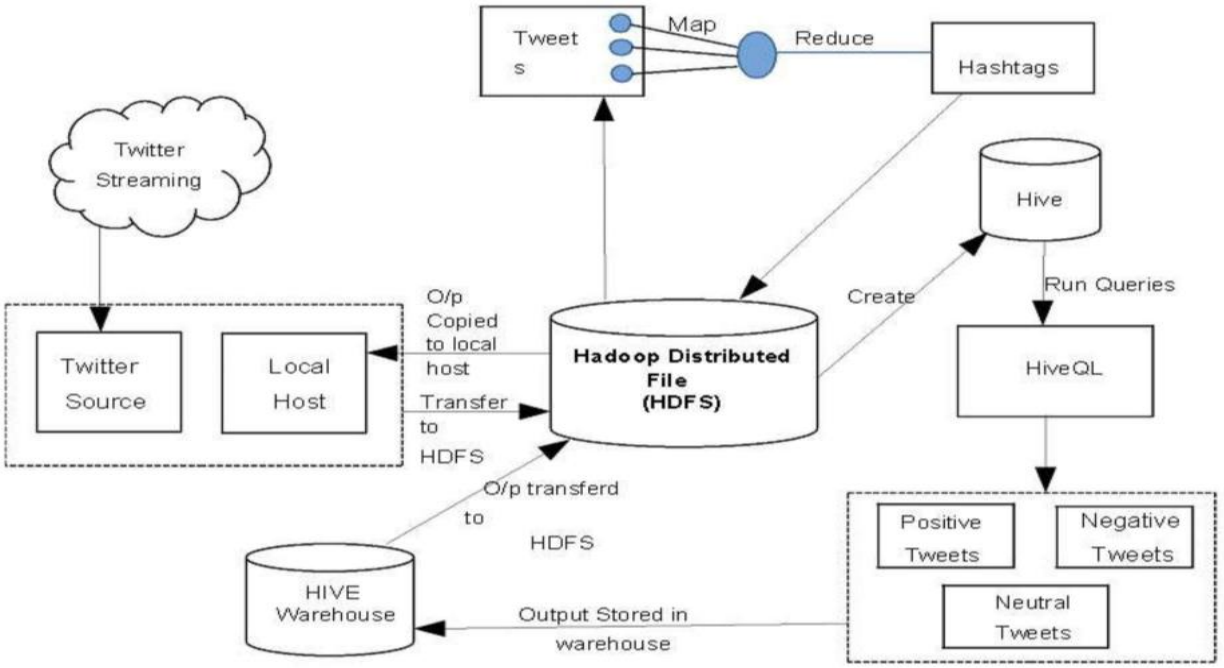
Total HDFS capacity per node	64GB
Number of nodes	4+
Total HDFS capacity on the cluster	256GB
Actual HDFS capacity (with replication)	43GB
/tmp mount point	20GB
Installation disk space requirement	12 GB

4.4 Internet band width

- 20Mbps internet speed is sufficient to fetch stream data from developer account.

3. DESCRIPTION

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 positive and negative 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 figure.



Election related tweets from twitter database will be the main dataset for the system. Thesaurus will be given for Sentiment analysis (positive/negative). To do our analysis we have used the popular big data tools in order to cope up with massive amount of text data. 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. 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]. 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. 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. 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.

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 Levenshtein 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 a 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 a 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.

5. RESULT:

Based on Hadoop platform and java for Map Reduce framework, we have implemented this system AS

Dataset

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

Results

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.

System

The OS used was Ubuntu- 16.04, with 4GB RAM and algorithms were implemented in JDK 1.8. Experiments were done on machine with core- i3 processor 2.0 GHz x 2.

```
SELECT text, word FROM twitter.tweets

LATERAL VIEW explode(split(text,' ')) text_ex as word;
```

The explode() is a Hive built-in User Defined Table-Generating Function (UDTF) that breaks down an array into its elements. In this case the tweet gets broken into words. The LATERAL VIEW joins the output of explode() to the input row (tweet) creating a result set that contains n rows (words) for each tweet.

We used following queries for collecting 12 extremely frequent hash-tags on the respective data.

```
SELECT LOWER(hashtags.text),

COUNT(*) AS total_count


FROM tweets

LATERAL VIEW  EXPLODE(entities.hashtags) t1

AS hashtags
```

GROUP BY LOWER(hashtags.text)

ORDER BY total_count

DESC LIMIT 14;

Result-on:

#SoniaGandhi 211

#Indian 142

#RahulGandhiInMumbai 139

#BJP 112

#Modi 106

#RahulGandhi 73

#Congress 60

#Kejriwal 40

#NarendraModi 38

#soniagandhi 29

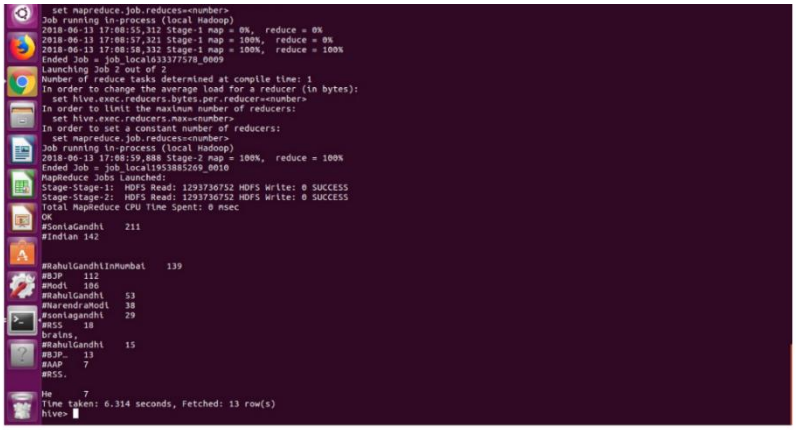
#RSS 18

#Congress 17

#AAP 17

#RSS 17

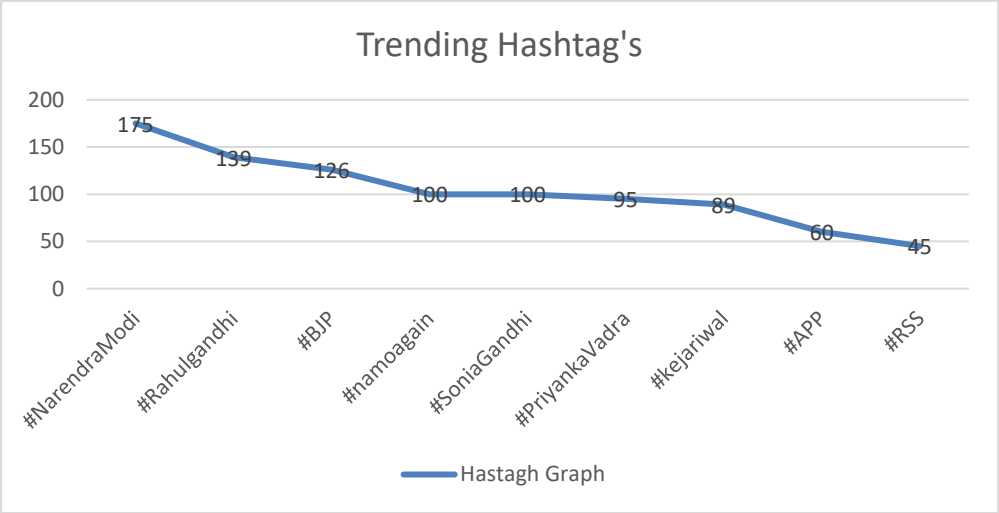
- **Hashtag wise tweet**



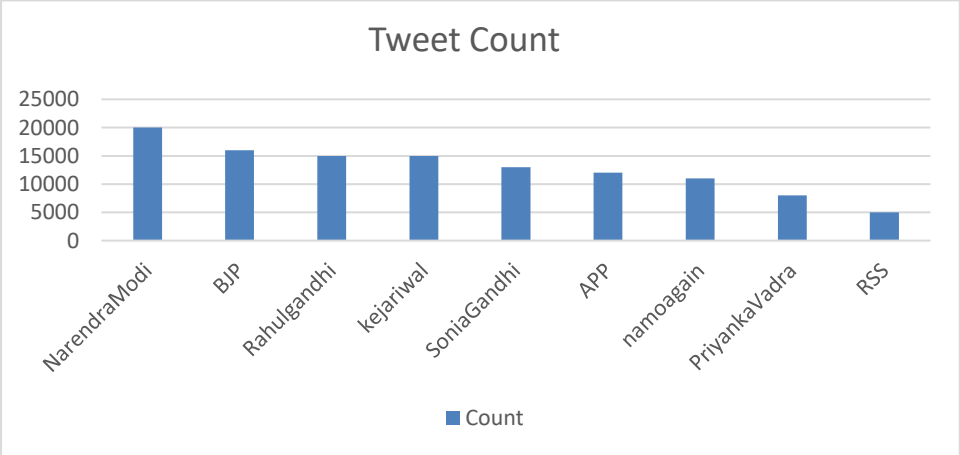
- **Hashtag wise tweet count**

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

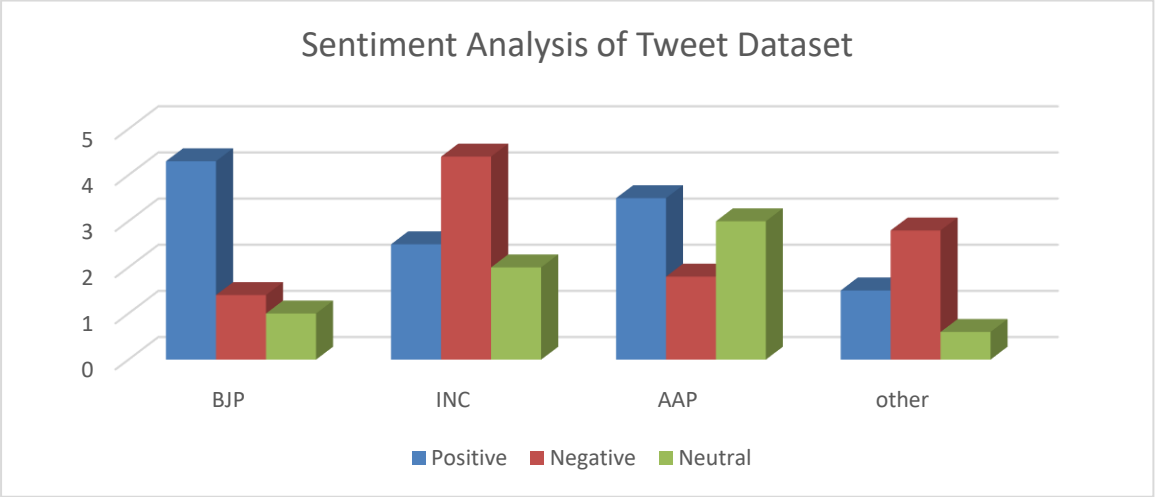
- **Trending topic of elections**



- **Count**



- **Sentiment Analysis of Tweet Dataset**



6. CONCLUSION:

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 will work 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.

7. REFERENCES:

[1] India is now world's second largest Internet user after China. https://www.statista.com/topics/2157/internet-usage-in-india/
[3] Gayatri Wani , Nilesh Alone, "A Survey on Impact of Social Media on Election System" http://www.ijcsit.com/docs/Volume%205/vol5issue06/ijcsit20140506100.pdf
[5] Social media for political campaign in India. http://www.slideshare.net/RaviTondak/social-media-for-political-campaign
[6] 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.
[7] AbhishekBhola "Twitter and Polls: Analyzing and estimating political orientation of Twitter users in India General Elections2014" arXiv:1406.5059 [cs.SI]
[8] IoannisKatakis, Nicolas Tsapatsoulis, Fernando Mendez, VasilikiTriga, and ConstantinosDjouvas "Social Voting Advice Applications - Denitions, Challenges,Datasets and Evaluation" IEEE TRANSACTION CYBERNETICS ,VOl. 44 No. 7

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

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

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

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