Election Analysis and Prediction Using Big Data Analytics

Omkar Sawant, Chintaman Taral, Roopak Garbhe
Students, Department Of Information Technology
Vidyalankar Institute of Technology,
Mumbai, India
omkar.sawant6@gmail.com

Deepali Nayak
Assistant Professor, Department. Of Information Technology
Vidyalankar Institute of Technology,
Mumbai, India
deepali.nayak@vit.edu.in

***Abstract*—** The aim of this paper is to propose an alternate way to conduct elections by highlighting the fundamental loopholes present in current system. The current election process considers the number of votes polled by a candidate as the sole parameter to choose the winner. Our proposal aims at highlighting the discrepancy between the most deserving candidate to win the election and the candidate who is predicted to win on the basis of his or her popularity. We will be choosing the deserved candidate by taking into account parameters such as the criminal records, educational qualifications, past social work, previous term record etc. We want the election process to not be a contest of popularity only. We aim to bring about a change in the election process to make it better and unprejudiced.

Keywords-*Election Analysis, Prediction, Hadoop Framework, Pig Programming, Text mining, Big Data Analytics.*

*****

I. INTRODUCTION

The current Election system that is being followed only considers the number of votes gained by the candidates to decide the winner ignoring all other vital aspects. Also, a significant section of the population chooses not to vote as they believe that the entire system is flawed. As a result, wrong candidates get selected for high & important positions in Government offices and other organizations. We believe that votes given to a candidate just on the basis of the spurious claims made during their Election campaigns without knowing his or her qualifications, previous records and eligibility are not justifiable Hence, we came up with an altogether innovative idea that exploits the power of Hadoop for an in-depth analysis of elections for overcoming the flaws present in the current system. Our proposed system for conducting elections is designed to increase people's participation in elections and to make it more stringent and impeccable.

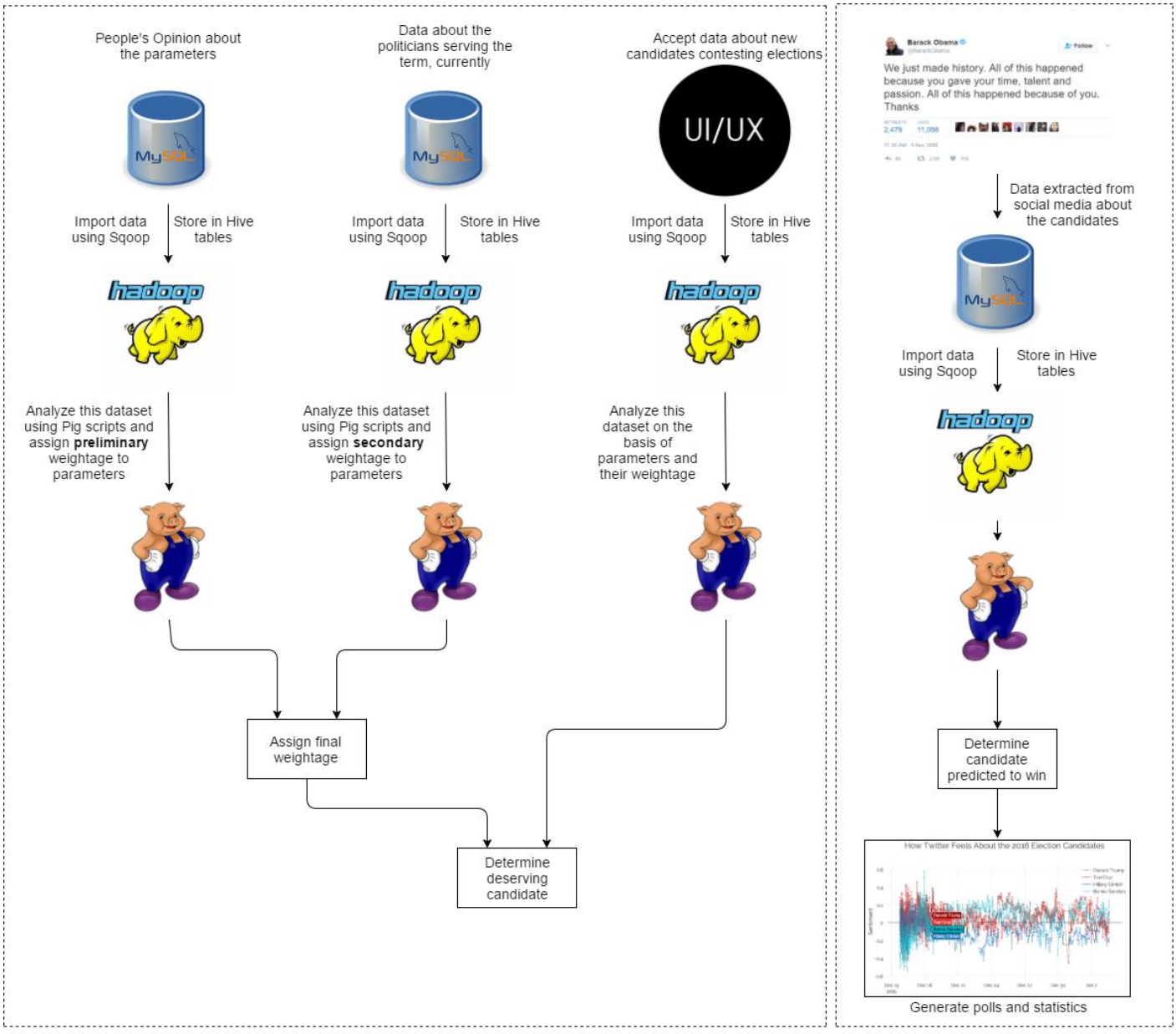
A. Proposed System

This project will provide a compact and lightweight distributed task execution framework that has been divided into two stages:

1: Analysis stage

2: Prediction stage

 ISSN: 2321-8169
107 – 111



B. Analysis stage

The detailed description of the analysis stage is as follows:

The cardinal goal of analysis stage is to determine the deserving candidate to win the election from the given list of candidates. The candidates will be judged according to following parameters:

- Educational Qualifications
- Criminal records
- Past social work
- Previous term record
- Speaking, motivational skills & personality
- Social status and popularity

To use the above parameters to judge the candidates it is important to give each parameter appropriate weightage. To do this we have performed analysis on two huge datasets,

1: Common people's opinion about these parameters. This data has been collected by conducting physical surveys as well as using Google Forms. Now, after analyzing this dataset, we have given a preliminary weightage to each of the parameters.

2: Data about the candidates who have or who are serving their term. This dataset contains various fields like tenure, number of useful projects done for benefit of people, track record, consistency, and data about the above mentioned parameters for each candidate etc. We conducted physical surveys and collected maximum data possible; however, as the political data is confidential, we were forced to make our own dataset. Hence, we formed our own data set by coding which consisted of data about thousands of candidates serving their term. Hence, after analyzing this dataset, we got a brief idea about the importance each of the parameter and a secondary

weightage was given to each of them irrespective of the weightage given in the first step.

The final weightage given to parameters is the average of both of them. This has drastically enhanced the accuracy of analysis as we are using public opinion as well as actual data. After we found out the final weightage about the parameters, we accepted data about the candidates presently contesting elections through a user interface and found out the deserving person among them by using the parameters. A score was generated for each candidate depending on his qualifications on each of the parameter and the result about the deserving candidate was stored in the database.

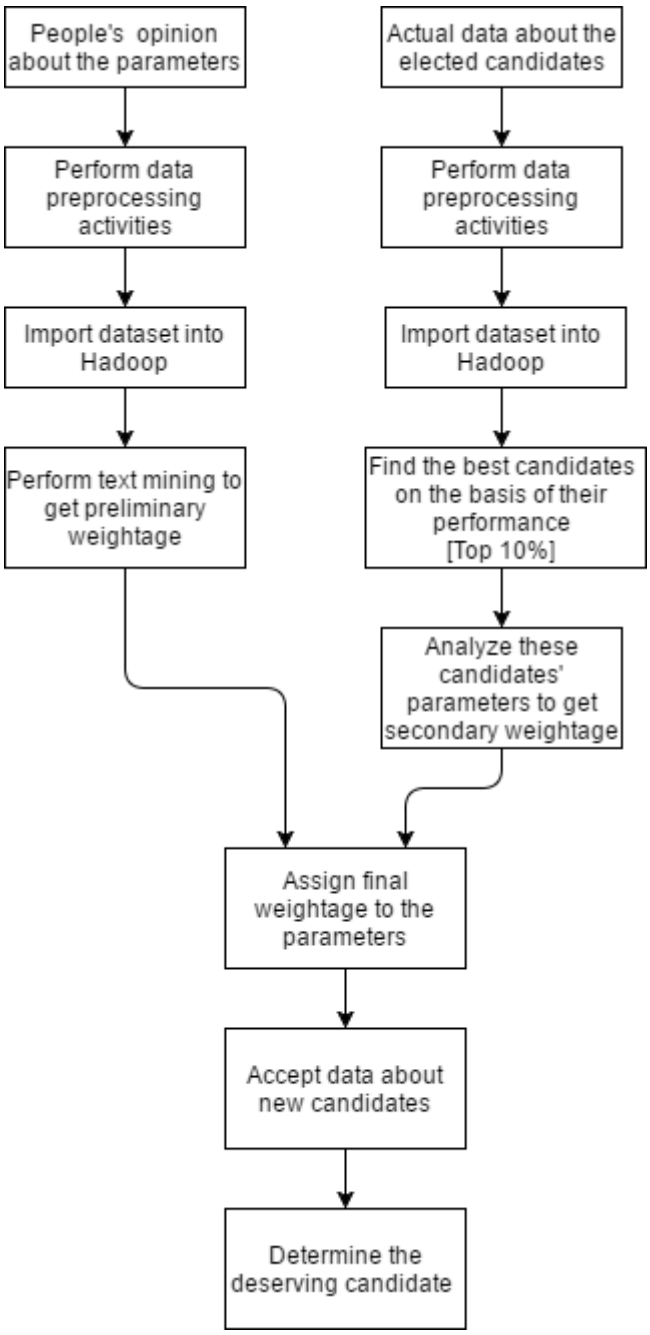


Figure 1. Flow diagram of Analysis phase

C. Prediction stage

The detailed description of the prediction stage is as follows:

• Data about the candidates (or their party) will be extracted out from social media [1].
E.g.: Twitter, Facebook, etc.
• Based on analysis of this data a candidate who has more chances of winning the election irrespective of whether he is deserves to win will be predicted.

This phase is very similar to the exit polls that are conducted during any major elections.

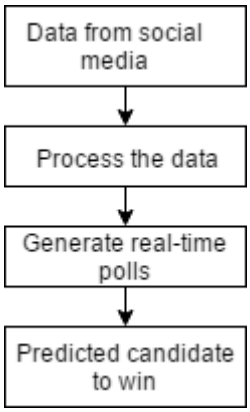


Figure 2. Flow diagram of Prediction phase

II. ALGORITHM/PSEUDO-CODE

Step 1: Conduct surveys and gather information about people's opinions about the parameters.

Step 2: Perform pre-processing activities & bring the data into suitable form for transferring it to Hadoop ecosystem.

Step 3: Using Sqoop tool, import the dataset containing people's views into Hive tables.

Step 4: Make appropriate set of positive, negative and neutral keywords that will be used for performing Text mining on the dataset.

Step 5: Using pig script, perform Text mining on the dataset and evaluate a preliminary weightage to each parameter.

Step 6: Store this result into Hadoop Distributed File System (HDFS) for future use.

Step 7: Collect data about the politicians currently serving or has served their term.

Step 8: Perform data cleaning & integration activities and add required fields useful for analysis.

Step 9: Import this data set using Sqoop into Hive tables.

Step 10: Using fields like total number of projects done by the candidate for people's benefit, clear track record and

overall satisfaction score find out the top 10% successful politicians.

Step 11: Use Pig programming to analyze the parameters of the top candidates to again give a secondary weightage to the parameters.

Step 12: Store this result into HDFS.

Step 13: Integrate the two weightages and give a final weightage to each of the parameter.

Step 14: Accept data about the candidates now contesting elections.

Step 15: Using the weightage, judge the candidates on the basis of the parameters and generate score for each candidate.

Step 16: The candidate with the highest score is the deserving candidate.

Step 17: Direct this output into a .txt file and add necessary explanations.

Step 18: Extract data from social media and other sources about the candidates contesting elections to have a brief idea about their popularity.

Step 19: Extract this data into Hive tables.

Step 20: Using pig programming, Batch processing and the collected data generate polls to predict the winner among those by creating graphs and charts.

Step 21: Direct this output to the same .txt file.

Step 22: The .txt file will show the difference between the deserved candidate & the candidate predicted to win thus highlighting the flaws of current system.

III. TECHNOLOGIES & CONCEPTS USED

A. *Hadoop Ecosystem*

As we started planning on our project, we realized that to do this kind of analysis, there will be a need for performing analysis on a huge amount of data. Moreover, the data will be available in an unstructured form as there is no predefined approach of collecting people's opinion in a uniform way. Hence, as we explored our options and decided to use Apache Hadoop as it is flexible and there is no need to pre-process data unlike Relational Databases. Also, Hadoop offers a convenient way to handle unstructured data and is an open source software platform for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware [2]. It also provided various tools like Hive, Pig, Sqoop that were perfectly suitable for our project.

B. *Apache Pig*

Apache Pig is a high-level platform for creating programs that run on Apache Hadoop [3]. As mentioned earlier, in the analysis phase, there is a need to analyze people's opinions about the parameters as well as actual data about the politicians in order to give those parameters a proper weightage. This we have achieved by writing Pig scripts. For scrutinizing people's opinions we have used Sentiment Analysis. Sentiment analysis is the process of using text analytics to mine various sources of data for opinions. Also for prediction phase, pig scripts & batch processing is used to generate polls. For real time & dynamic prediction analysis the Pig Scripts can be replaced by Apache Spark.

C. *Apache Hive*

The Apache Hive provides data warehouse software which facilitates reading, writing and managing large datasets residing in distributed storage using SQL. The data that we collected by conducting physical surveys and online has been imported into Hive tables using Sqoop. Sqoop is a tool which is used to transfer data between Hadoop Distributed File System (HDFS) and external databases and data warehouses. We have used it as a repository to store the datasets Also, we have fired ad hoc queries on the data stored in Hive tables in order to create statistics, graphs and charts.

IV. CONCLUSIONS

The discrepancy and the loopholes in the present system are clearly highlighted with the help of this innovative project. After examining the result, it was clear that the current election process is unable to select the most suitable candidate of the available people. Hence, pertaining to the current quality of politicians, our proposed system should be implemented by the Government as early as possible by making required changes. This type of analytics based approach can be used for any type of election.

This approach also highlights the power and scope of data analytics as a whole in the Computer Science field.

REFERENCES

[1] A Review of Sentiment Analysis in Twitter Data Using Hadoop (paper published in International Journal of Database Theory and Application)J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Dhruba Borthakur, "The Hadoop Distributed File System: Architecture and Design", The apache Software Foundation, 2007.K. Elissa, “Title of paper if known,” unpublished.

[3] Processing performance on Apache Pig, Apache Hive and MySQL cluster. (Date Added to IEEE Xplore: 15 January 2015 INSPEC Accession Number: 14853176)