Volume 5, Issue 4, April 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Data Analysis on Current Affairs Using R

Manjunath Mulimani, Nireeksha V Shetty, Nikshitha Shetty, Renita Maria Lolo, Rajashree L
Dept. Of Computer Science, Sahyadri College of Engineering & Management,
Karnataka, India

Abstract--People react to various events, topics and entities of the world by expressing their personal views and emotions. Peoples' opinion can be known from newspapers, blogs and websites on recent happenings. These sources help us collect massive amount of data and give an insight about Big-data and its application in Indian-political scenarios. Emotions expressed usually vary from one person to another. These emotions can correspond to a wide range of intensities that vary from very mild to strong. Analyzing the emotions require an adequate processing and understanding of these expressions. In the recent years these kind of analysis have become a main part of several research fields. Business, marketing and politics are some of these fields. We present a system that gives an overview of frequently discussed subjects. Our main aim has been to highlight the extraordinary potential of big-data application on current political activities of the nation.

Keywords: Big data; Python; R; Sentiment

I. INTRODUCTION

Large complex datasets require an enormous amount of computing power to work with is referred as big data [1]. Big data not only has a great potential to change and cause a revolution in research, but also the way big corporations governments and organizations run business and find patterns. This influences the companies the way they make decisions. Humans are influenced by the environment and external factors they are surrounded with. They communicate by expressing their feelings, opinions and preferences they deal with in various ways [2]. It is very important to know the emotional load of a message, expressed in either written or verbal, when it comes to understanding its true meaning. Therefore, opinion and sentiment play a key role in human interaction. A number of studies have been evolved over the years individually and also collectively, in order to understand the human behavior. The collection or analysis of opinions and sentiment correspond to the need of measuring the impact that an entity has on a group of individuals. Various studies have emerged in order to understand the social sentiment in politics and forecast election outcomes, and also in business, to predict the growth and success of a certain product so that it can be recommended to others.

News can be good or bad based on the people's mindset. Although full comprehension of natural language text remains well beyond the power of systems, the statistical analysis of sentiment can provide a surprisingly meaningful sense of how the entities are affected by the latest news [3]. In the early days social media was not as popular as of today and there was no means of collecting large set of data. Gathering data on opinions was usually achieved at very small scale and was very expensive. As the web started getting popular the communication gradually increased making the social networks overloaded with opinionated data. Due to this social media has widened new possibilities for human interaction. Nowadays micro blogging platforms provide real-time sharing of comments and opinions.

This paper is organized as follows: Section II describes the related works. Section III presents the system architecture. Section IV describes methodology. Section V involves the data analysis and results. Section VI gives the conclusion.

II. RELATED WORK

Automatic sentiment analysis of Twitter messages is donewith automatic training based in tweets containing either emoticons or sentiment-based words. These sets were used to categorize the tweets that could not be classified automatically .Many approach were proposed. First uses the sentiment incorporated in the emoticons as a criterion to automatically classify the messages. Second uses words that express sentiment as criteria for the automatic classification

The tweet sentiment analysis is not only based on text content alone, such as lexicon based classifiers, but also by combining Natural Language Processing and Machine Learning techniques to use both content as well as connectivity patterns between Twitter users.

A lot of existing work has been done using dictionaries that capture the sentiment of words on sentiment analysis based on content. Xujuan Zhou [4] proposed a Tweets Sentiment Analysis Model (TSAM) that can spot the societal interest and general people's opinions in regard to a social event. Jalaj S.Modha [5] proposed Automatic Sentiment Analysis for Unstructured Data which classify and handle subjective as well as objective statements for sentimental analysis which is applied on closed domain. (Indian political news articles).

Keke Cai [6] proposed a novel topic detection method using point-wise mutual information and term frequency distribution. Other techniques could detect the topics that are highly correlated with the positive and negative opinions.

Such techniques, when coupled with sentiment classification, can help the business analysts to understand both the overall sentiment scope as well as the drivers behind the sentiment.

Kowcika proposed a system which is able to collect useful information from the twitter website and efficiently perform sentiment analysis of tweets [7]. The system uses efficient scoring system for predicting the sentiment of the tweets. Then, Sentiment Classifier Model labels the tweet with a sentiment.

III. ARCHITECTURE

The data for this study was obtained from website www.pmindia.gov.in/en/, www.newshunt.com, www.narendramodi.in/and various news papers.These websites consist of a huge source of data of government programs and nation's issues. As part of this project, a Scrapper is written to collect transcripts of modi speeches and remarks. These transcripts are written in html format including images.The dataset for this study is created by analysing massive amount of html data and measuring its various relevant features and hidden patterns.

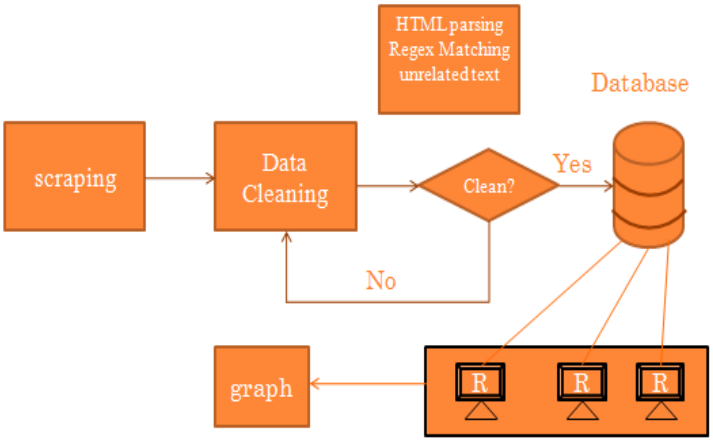

Fig.1 System Architecture

IV. METHODOLOGY

We will now move on to look at the different stages and procedures that helped us pulling out insightful results, from gathering relevant data to storing and analysing it.

A. Scraping

The format of the transcripts is encoded in html.The technique in which a computer program extracts data from human-readable output coming from websites is referred as data scraping. We built a scraper to gather required data from different website, this scrapper looks for various URLpattern and stores its html content. Web Scraping is a process of extracting the html data from a Web page. Since the scraped data is embedded within html tags, it is necessary to remove the tags as we are interested only in the contents enclosed within the tags [8]. Without clean data, every big data initiative will take longer, cost more, and deliver fewer benefits.Data preparation is essential in order to assure the data is useful and functional toward the intended end analysis web page to another format.

B. Data Preparation

1) Extracting the important data

Data extraction is the process of getting data out of data sources for further processing or storage of data [9]. Typical data sources include documents, web pages,PDF,emails etc. It has been a big challenge extracting data from these unstructured sources. The majority of extraction is done from these unstructured data sources.

We extracted the contents within html <p> tags as shown in the Fig 2:

```
<p><strong>PM chairs meeting of National Ganga River Basin authority</p><p>The Prime Minister, Shri Narendra Modi, today called for an meeting</p>
```

Fig 2: Html Page Contents

2)Clean the imported data

Once the required data has been extracted it should undergo cleaning process**.**

```
Soup=BeautifulSoup("http://www.narendramodi.in")

Res=Soup.select("p")

Print Res
```

We made use of BeautifulSoup to extract the contents of <p> tag. BeautifulSoup is a python library for pulling data out of html and xml files[10].Regular expression methods were used to remove the special characters,whitespaces etc.We also removed English stopwords from the data.Every cleaned transcript was stored in database for further processing.

V. DATA ANALYSIS AND RESULTS

Our main intention in this paper is to analyse the governance of modi government from May 2014.For this purpose we make use of R.R framework provides a means to analyze the sentiment. Sentiment analysis is done using "sentiment" package.

It includes two handy functions:

classify_emotion-Classifies the emotion based on naive Bayes classifier.These emotions are classified into six different categories such as anger, joy, sadness, disgust, surprise, fear.

classify_polarity- Classifies the polarity based on naive Bayes classifier.The polarity can be positive or negative.We analyze the following attributes [2].

A. Polarity

This aims to extract polarity information from a passage.Usually positive, negative and neutral values are obtained from polarity-oriented method. The polarity-oriented lexical resources contain list of positive and negative words as shown in Fig 5.

B. Strength

According to the polarity sentiment dimension it provides the different intensity levels.The scores indicate the strength of sentiments (positive or negative)expressed in a document.These numerical scores are obtained from the strength-oriented methods. Strength oriented lexical resources provide the intensity scores which describes the positivity and negativity

C. Emotion

It focuses on extracting emotion or mood states from a document. An emotion-oriented method classifies the message into different categories.Some of the emotional categories are sadness, joy, surprise, etc...Expressions are marked according to different emotion states and are provided by emotion-oriented lexical resource as shown in Fig 6 and Fig 7.

To work in distributed systems pbdR (Programming with big data in R) is used.

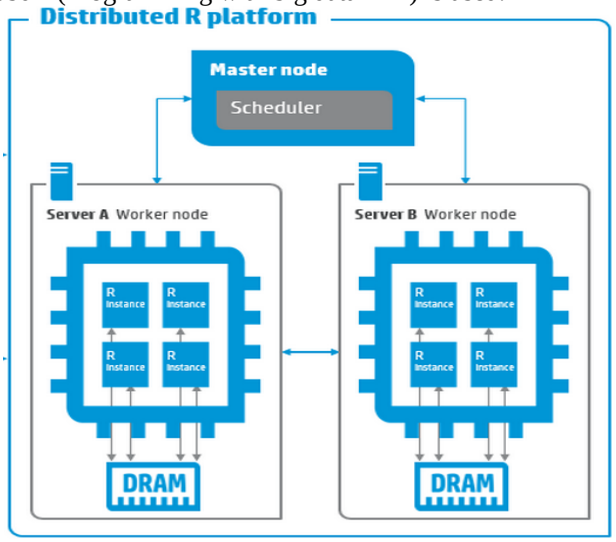

Fig 3: Parallel Analysis in R

Positivity, negativity is represented in terms of graphs and wordcloud as shown in Fig 7

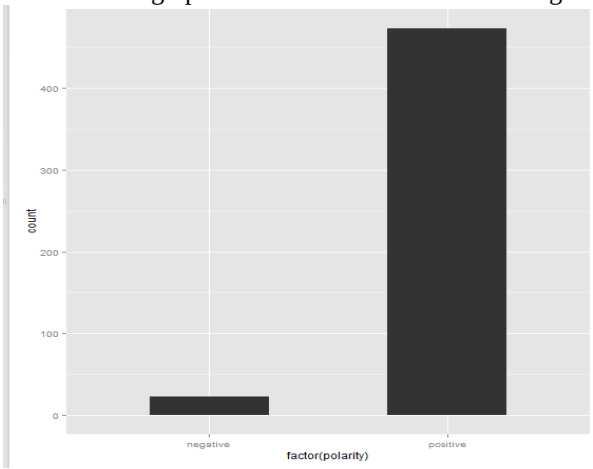

Fig. 4 Positive and negative count

Console D:/mod/

address	address	4	unknown	positive
also	also	4	unknown	positive
around	around	4	unknown	positive
country	country	4	unknown	positive
first	first	4	unknown	positive
global	global	4	unknown	positive
hannover	hannover	4	unknown	positive
indo	indo	4	unknown	positive
leaders	leaders	4	unknown	positive
like	like	4	joy	positive
open	open	4	unknown	positive
permanent	permanent	4	unknown	positive
seat	seat	4	unknown	positive
war	war	4	anger	negative
year	year	4	unknown	positive
000	000	3	unknown	positive
body	body	3	unknown	positive

Fig. 5 Polarity, Emotion and Frequency Count

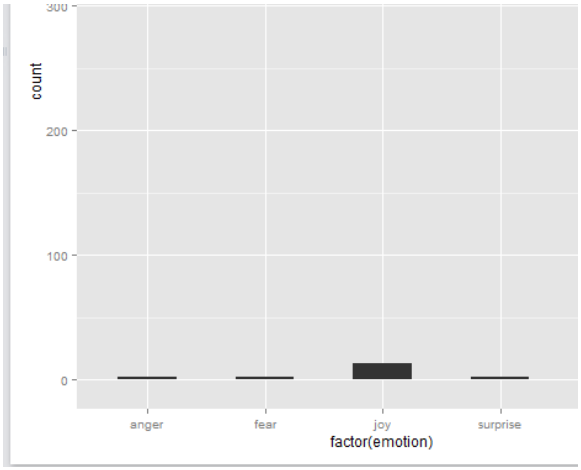

Fig. 6 Emotion

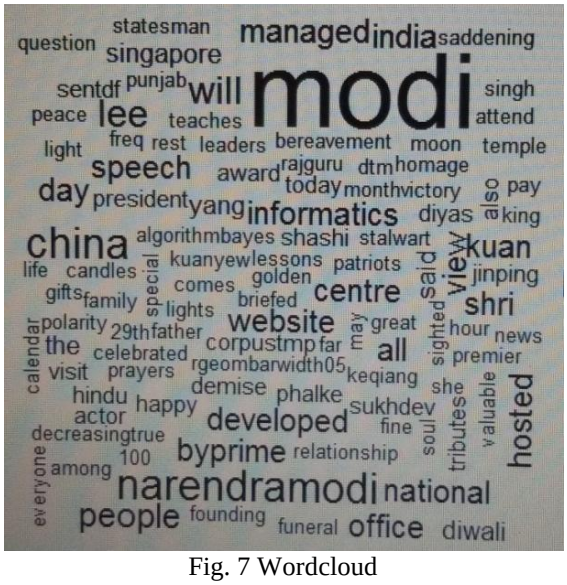

Fig. 7 Wordcloud

Our aim of this research was to find a way to answer these questions:
1. What were the most referenced regions, areas?
2. What were the most spoken or referred words used by Modi?
3. Was there any big change in speech patterns within Prime ministerial debate period?
4. What was Modi's focus, Internal or foreign affairs?
5. Which were the less/more concentrated subjects by him?

To answer those questions, we needed to find relevant parameters to measure:

A. Words Frequency

This parameter helps to find the most referred subjects by modi over different time period [1].

B. Sentiment Analysis

Sentiment analysis is one of the important types of text analysis. It aims to make decisions by extracting and analysing opinions expressed in the form of texts. It identifies positive and negative opinions and measures how positive or negative an entity is regarded.

VI. CONCLUSION

This article has attempted to tackle two major tasks. Firstly, it has highlighted benefits of Big-data to Analyse and measure political events. Secondly it provides an overview on the modi govt, opinions of modi on different scenarios. Our main aim has been to showcase the extraordinary potential of Big-data application on current political activities.

REFERENCES

[1] Anass Bensrhir - *Big Data for Geo-political Analysis application on Barack Obama's remarks and speeches*Published in Computer Science and Applications (AICCSA), 2013, ACS International Conference

[2] Felipe Bravo Marque, Marcelo Mendoza, Barbara Poblete- *Meta-Level Sentiment Models for Big Social Data Analysis*

[3] Namrata Godbol, Manjunath Srinivasaiah, Steven Skiena- *LargeScale Sentiment Analysis for News and Blogs*

[4] Xujuan Zhou, Xiaohui Tao, Jianming Yong , Zhenyu Yang-*Sentiment analysis on tweets for social events* Published in: Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference, 557 - 562 IEEE 27-29 June 2013.

[5] Jalaj S .Modha, Gayatri S.Pandi, Sandip J .Modha-*Automatic Sentiment Analysis for Unstructured Data*Volume 3, December 2013

[6] Keke Cai, Spangler S ,Ying Chen , Li Zhang-*Leveraging Sentiment Analysis for Topic Detection*," Web Intelligence and Intelligent Agent Technology 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on (Volume:1) IEEE 9-12 Dec. 2008

[7] A Kowcika, Aditi Gupta, Karthik Sondhi, Nishit Shivhre Raunaq Kumar-*Sentiment Analysis for Social Media*Volume-3 July 2013

[8] *http://www.webopedia.com/TERM/W/Web_Scraping.html*

[9] *http://en.wikipedia.org/wiki/Data_extraction*

[10] *www.pythonforbeginners.com/beautifulsoup/*

