SENTIMENTAL ANALYSIS OF YOUTUBE VIDEO COMMENTS

A Project Report Submitted
in Partial Fulfilment of the Requirements
for the Degree of

B.TECH(Hons)

in

COMPUTER SCIENCE AND ENGINEERING

by

Sai Siddardha Sumala
(Roll No. 2015BCS0032)

to

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY KOTTAYAM - 686635, INDIA
April 2019

DECLARATION

I, **Sai Siddarhda Sumala** (**Roll No: 2015BCS0032**), hereby declare that, this report entitled **‘SENTIMENTAL ANALYSIS OF YOUTUBE VIDEO COMMENTS’** submitted to Indian Institute of Information Technology Kottayam towards partial requirement of **Bachelor of Technology(Hon)** in **Computer Science And Engineering** is an original work carried out by me under the supervision of **Dr.Shajulin Benedict** and has not formed the basis for the award of any degree or diploma, in this or any other institution or university. I have sincerely tried to uphold the academic ethics and honesty. Whenever an external information or statement or result is used then, that have been duly acknowledged and cited.

Kottayam - 686635 **Sai Siddardha**

April 2019

CERTIFICATE

This is to certify that the work contained in this project report entitled **"SENTIMENTAL ANALYSIS OF YOUTUBE VIDEO COMMENTS"** submitted by **Sai Siddardha Sumala** (**Roll No: 2015BCS0032**) to Indian Institute of Information Technology Kottayam towards partial requirement of **Bachelor of Technology(Hon)** in **Indian Institute Of Information Technology Kottayam** has been carried out by him under my supervision and that it has not been submitted elsewhere for the award of any degree.

Kottayam - 686635
April 2019

Dr. Shajulin Benedict
Project Supervisor

ABSTRACT

Nowadays, Big Data And Data Mining Have Attracted A Great Deal Of Attention In The Information Industry, Due To The Wide Availability Of Huge Amounts Of Data And The Urgent Need For Turning Such Data Into Useful Knowledge Through Predictive Models.Corporate Companies Are Using Social Media For Improving Their Businesses, The Data Mining And Analysis Are Very Important In These Days.User comments are the most popular but also extremely controversial form of communication on YouTube. Their public image is very poor; users generally expect that most comments will be of little value or even in thoroughly bad taste. Nevertheless, heaps of comments continue to be posted every day. This Project Deals With Analysis Of YouTube Data. The Analysis Is Done Using Users Sentiments Features Such As Views, Comments, Likes, And Dislikes.

Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction

Millions of people are using social network sites to express their emotions, opinion and disclose about their daily lives. However, people write anything such as social activities or any comment on products. Through the online communities provide an interactive forum where consumers inform and influence others.Moreover, social media provides an opportunity for business that giving a platform to connect with their customers such as social media to advertise or speak directly to customers for connecting with customers perspective of products and services. In contrast, consumers have all the power when it comes to what consumers want to see and how consumers respond. With this, the companys success failure is publicly shared and end up with word of mouth. However, the social network can change the behavior and decision making of consumers, for example, mentions that 87 percent of internet users are influenced in their purchase and decision by customers review. So that, if organization can catch up faster on what their customers think, it would be more beneficial to organize to react on time and come up with a good strategy to compete their competitors.

Sentiment Analysis helps in determining how a certain individual or group responds to a specific thing or a topic. Usually, surveys are conducted to collect data and do statistical analysis.For Video Sharing, YouTube Is The Most Popular Site On The Web. According To A Recent Study, Youtube2 Accounts For 20 percent Of Web Traffic And 10 percent Of Total Internet Traffic. YouTube Provides Many Social Mechanisms To Judge User Opinion And Views About A Video By Means Of Voting, Rating, Favorites, Sharing And Negative comments etc

1.2 Sentiment Analysis Benefits

Sentiment analysis is used to understand market or crowd sentiments. Social media sites like twitter,Youtube, facebook etc feeds are an important source of sentiment analysis. It could be used in any place where it is important to understand mass sentiments, such as

1.2.1 Political Campaign

To understand how to steer political campaign, what sort of issues would touch peoples hearts and which one wouldnt be ineffective or would have -ve impact etc.

1.2.2 Product/Services launch

To understand how the market response is to a particular product or service launch, Which particular aspect/feature is being liked by the users and where a change is required

1.2.3 Sports

Sports is very close to peoples heart and drive a lot of downstream revenue activities. It is important to understand crowd sentiments for sports and accordingly change sports strategy or advertising strategy etc.

1.3 Sentiment Analysis Challenges

The biggest challenge to making an ”accurate” sentiment analysis tool is cutting through the traditional ways of framing the problem. The truth is that almost all sentiment analysis engines today are fundamentally wrong approaches.

The most common approach is what's called a ”lexicon-based” approach. It involves taking a huge number of words and tagging them with a sentiment score. This approach is fundamentally limited and cannot recognize complex phrasing, sarcasm, idioms, slang, etc. When you're dealing with accuracies around 80 percent here (very typical) it's hard to get any kind of clear signal.

The second piece is moving beyond the positive/neutral/negative framing of the problem. What matters isn't just that something is positive or negative, at a bare minimum we need to know just how negative something is. For instance, looking at: ”The battery life is only mediocre” and ”This is the worst phone I've ever purchased"Both are clearly negative, but the second phrase is FAR more negative than the first.

The third piece is noticing trends, very much like other people have mentioned, it's not just about overall sentiment, you want combinations of sentiment, keywords, named entities. Don't just tell me that people are upset, but show me what they're upset about!

Chapter 2

PRELIMINARY WORK

2.1 Algorithm Used

At first the user will input the data in web service module after that based on number of rows in data it will split data either like as small data or big data. If the data is big data the user will choose any graphs either like composition charts or comparison charts or distribution charts or relationship charts after that it will give a output graph and it will show it in web module.

2.2 Proposed Architecture and Taxonomy

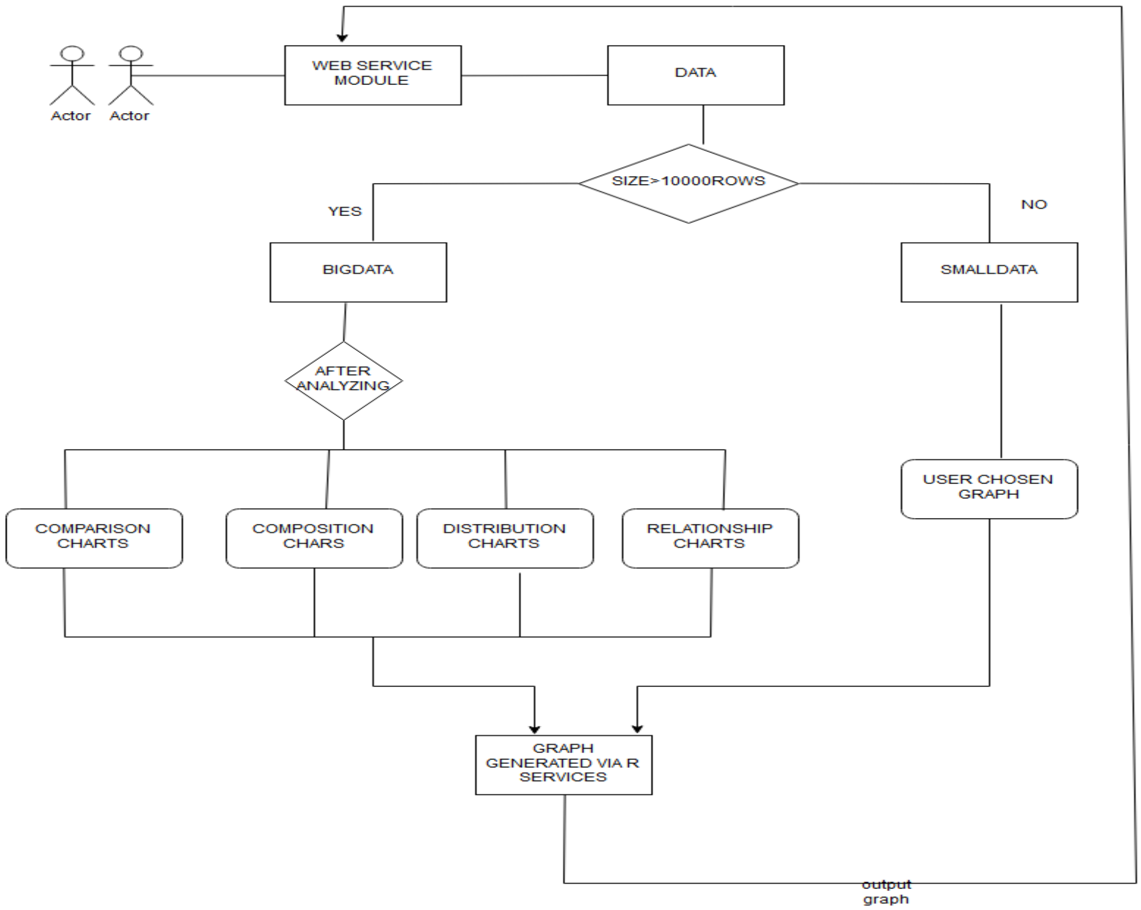

Figure 2.1: Proposed Architecture

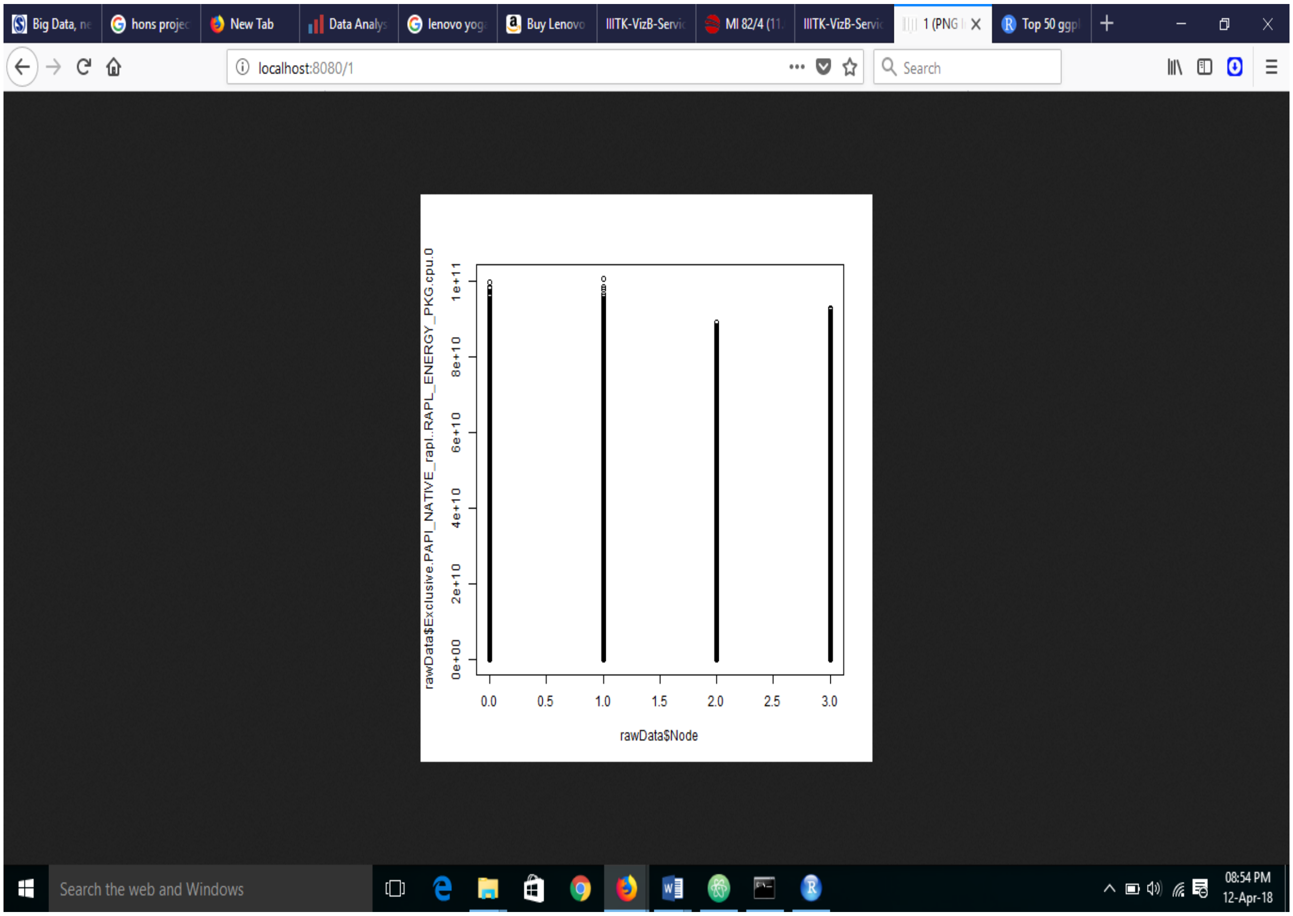

Figure 2.2: Scatter Chart

2.3 Performance Analysis of Algorithm

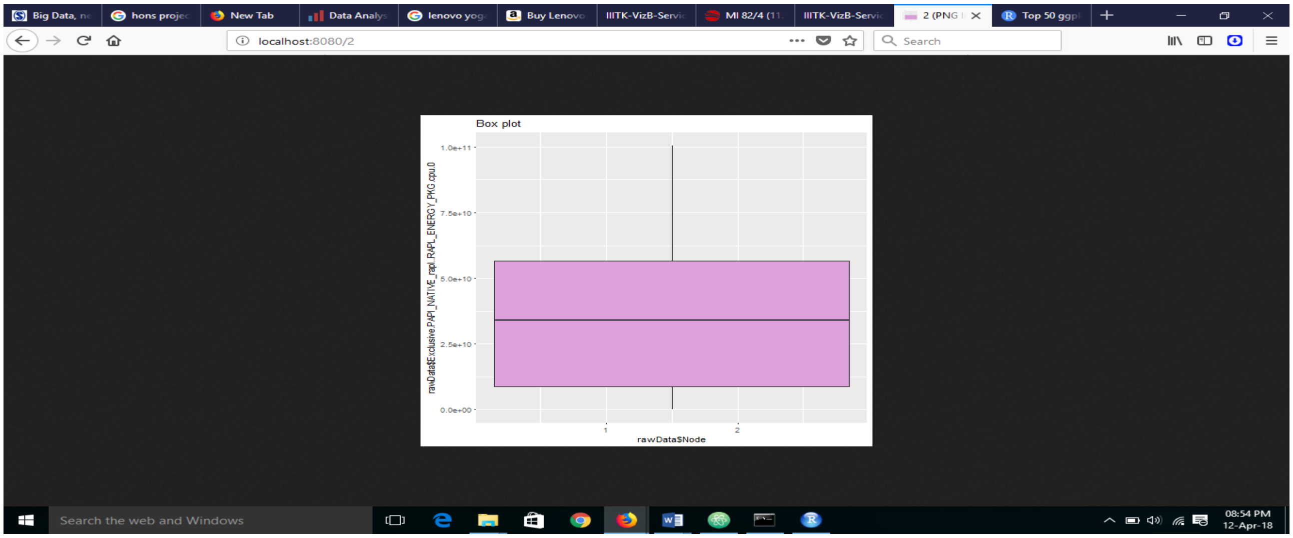

Figure 2.3: Box Chart

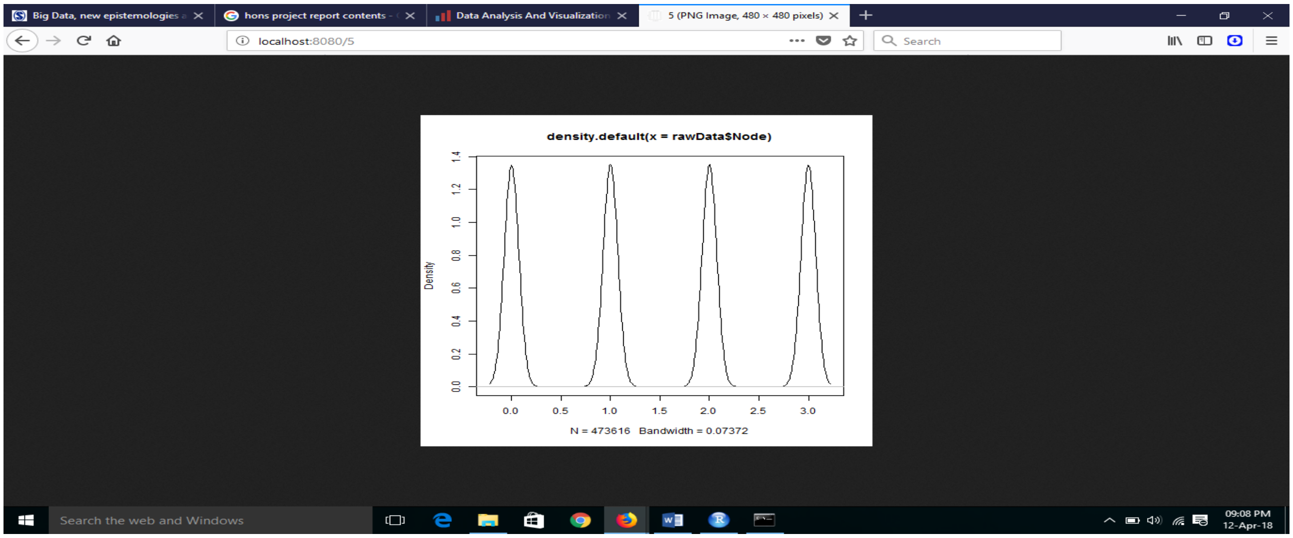
Big Data, new epistemologies a
hons project report contents -
Data Analysis And Visualization
5 (PNG Image, 480 × 480 pixels)
localhost:8080/5
Search
density.default(x = rawData$Node)
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0.0
0.5
1.0
1.5
2.0
2.5
3.0
N = 473616   Bandwidth = 0.07372
Search the web and Windows
09:08 PM
12-Apr-18

Figure 2.4: KERNALDENSITY chart for bigdata

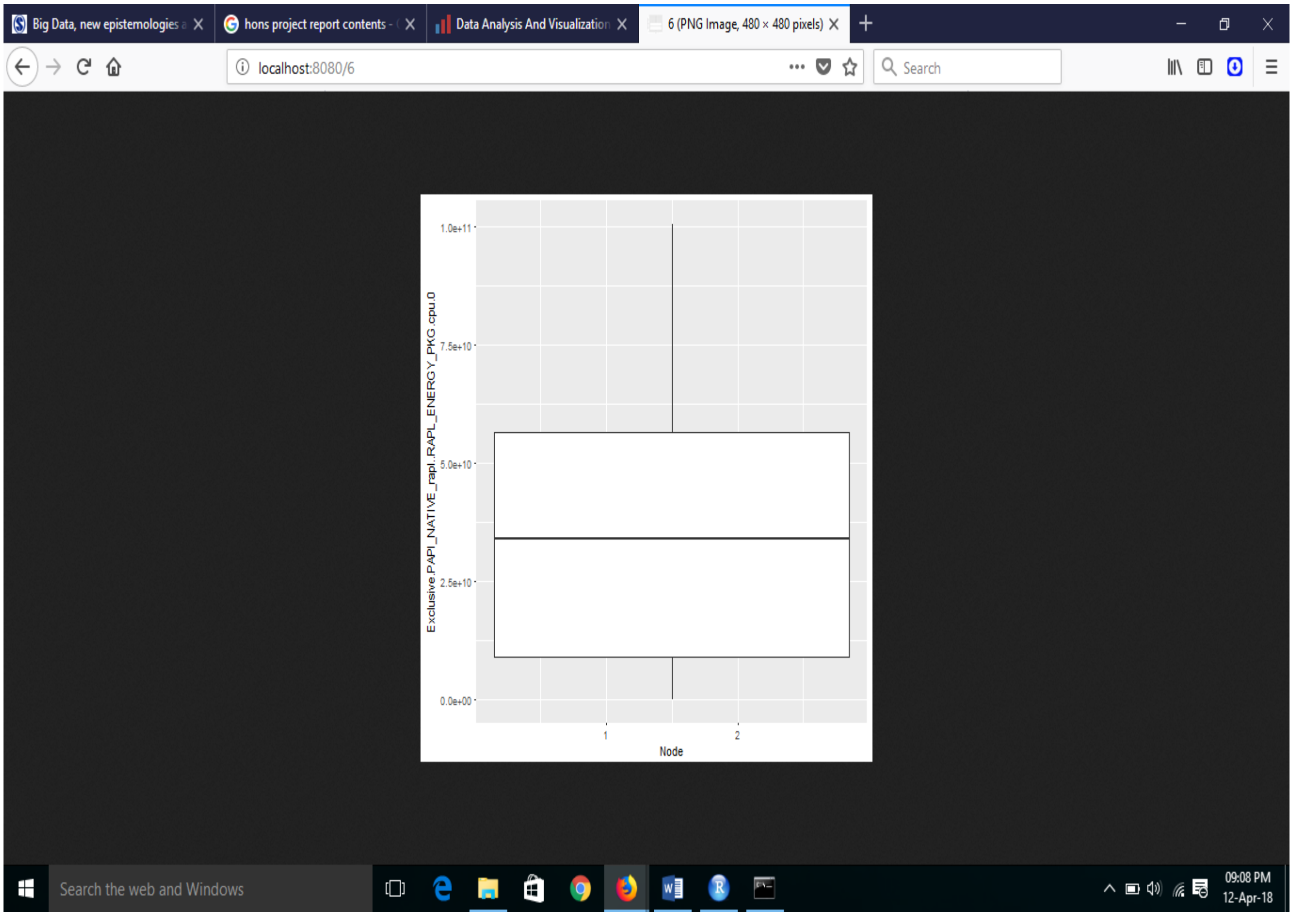
Big Data, new epistemologies a
hons project report contents -
Data Analysis And Visualization
6 (PNG Image, 480 × 480 pixels)
localhost:8080/6
Search
1.0e+11
7.5e+10
5.0e+10
2.5e+10
0.0e+00
Exclusive.PAPI_NATIVE_rapl..RAPL_ENERGY_PKG.cpu.0
1
2
Node
Search the web and Windows
09:08 PM
12-Apr-18

Figure 2.5: NOTCHEDBOX chart for bigdata

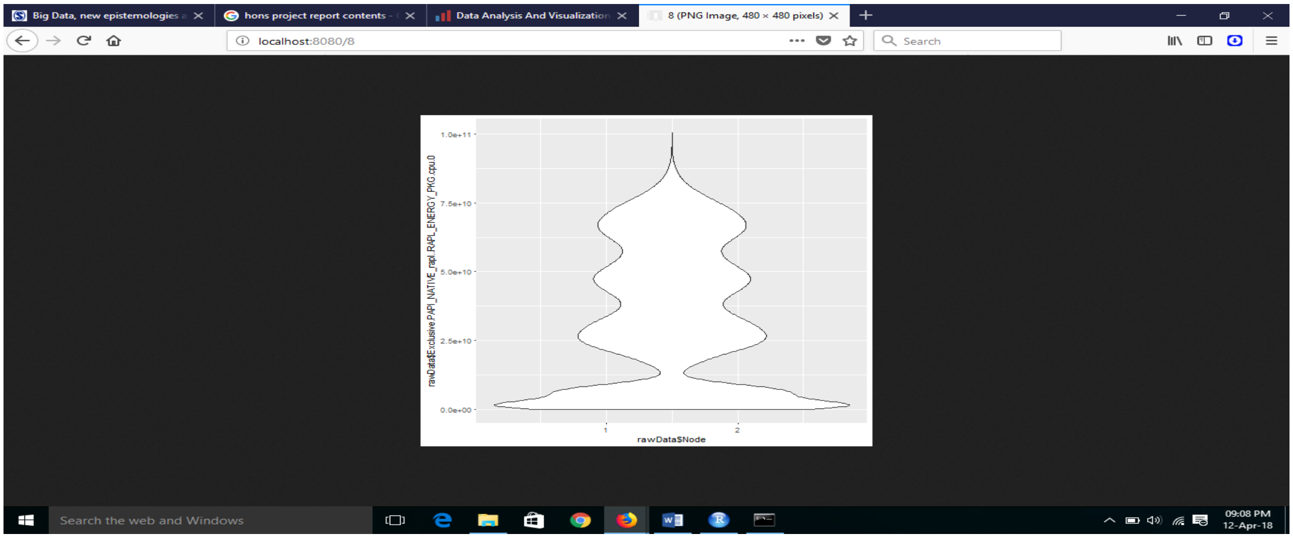
Big Data, new epistemologies a
hons project report contents
Data Analysis And Visualization
8 (PNG Image, 480 × 480 pixels)
localhost:8080/8
Search
1.0e+11
7.5e+10
5.0e+10
2.5e+10
0.0e+00
rawData$Exclusive.PAPI_NATIVE_rapl..RAPL_ENERGY_PKG.cpu.0
1
2
rawData$Node
Search the web and Windows
09:08 PM
12-Apr-18

Figure 2.6: VIOLIN chart for bigdata

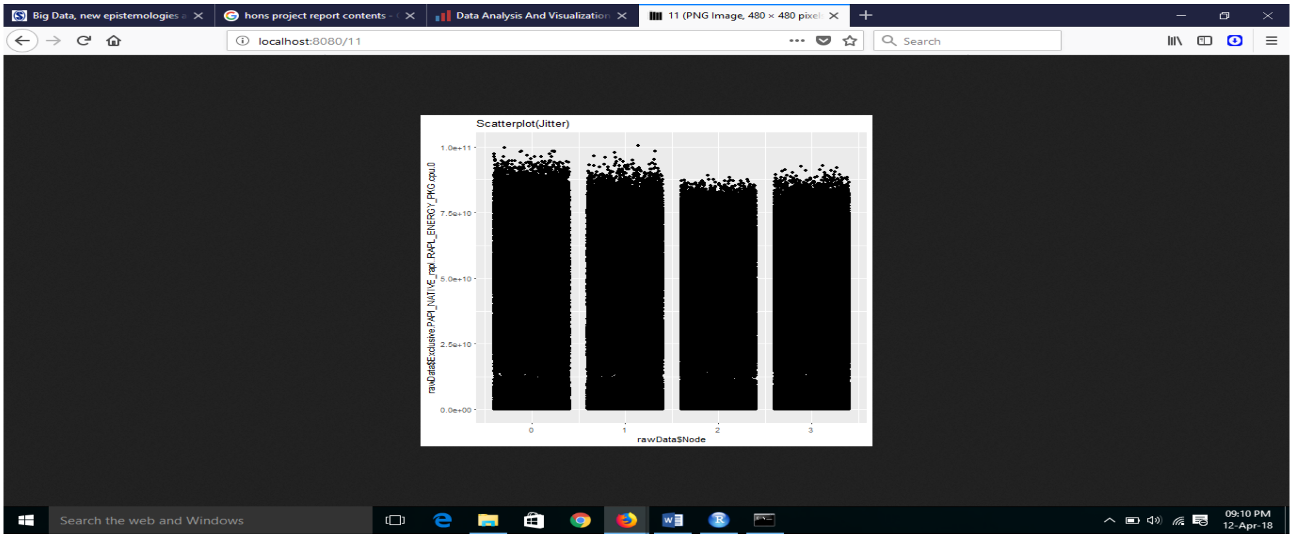
Big Data, new epistemologies a
hons project report contents
Data Analysis And Visualization
11 (PNG Image, 480 × 480 pixel
localhost:8080/11
Search
Scatterplot(Jitter)
1.0e+11
7.5e+10
5.0e+10
2.5e+10
0.0e+00
rawData$Exclusive.PAPI_NATIVE_rapl..RAPL_ENERGY_PKG.cpu.0
0
1
2
3
rawData$Node
Search the web and Windows
09:10 PM
12-Apr-18

Figure 2.7: JITTER chart for bigdata

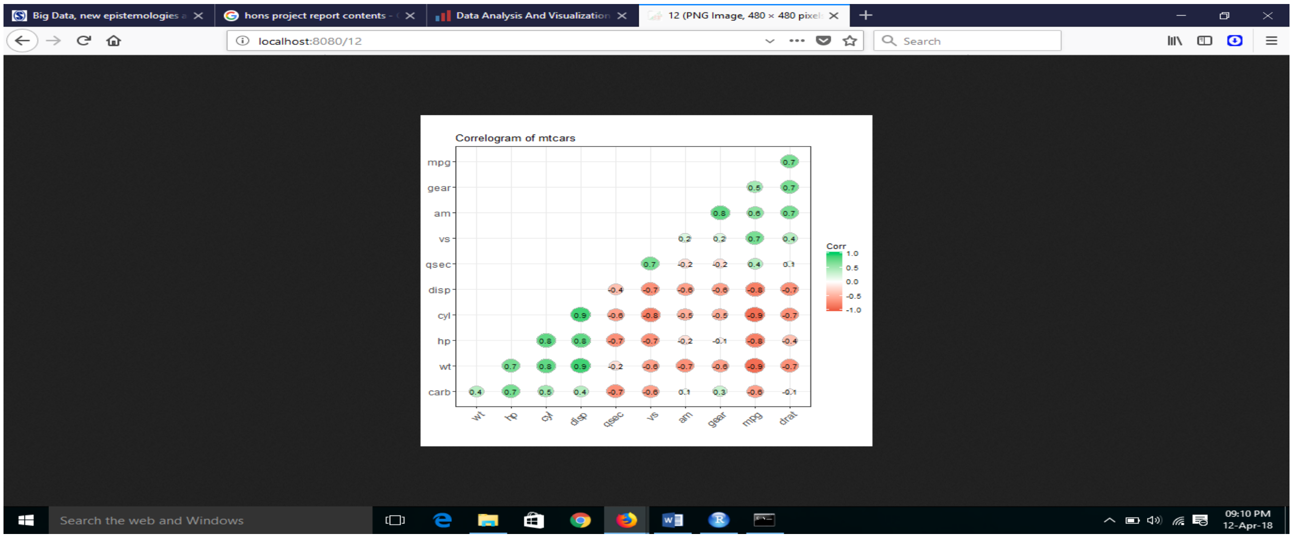

Figure 2.8: CORRELOGRAM chart for bigdata

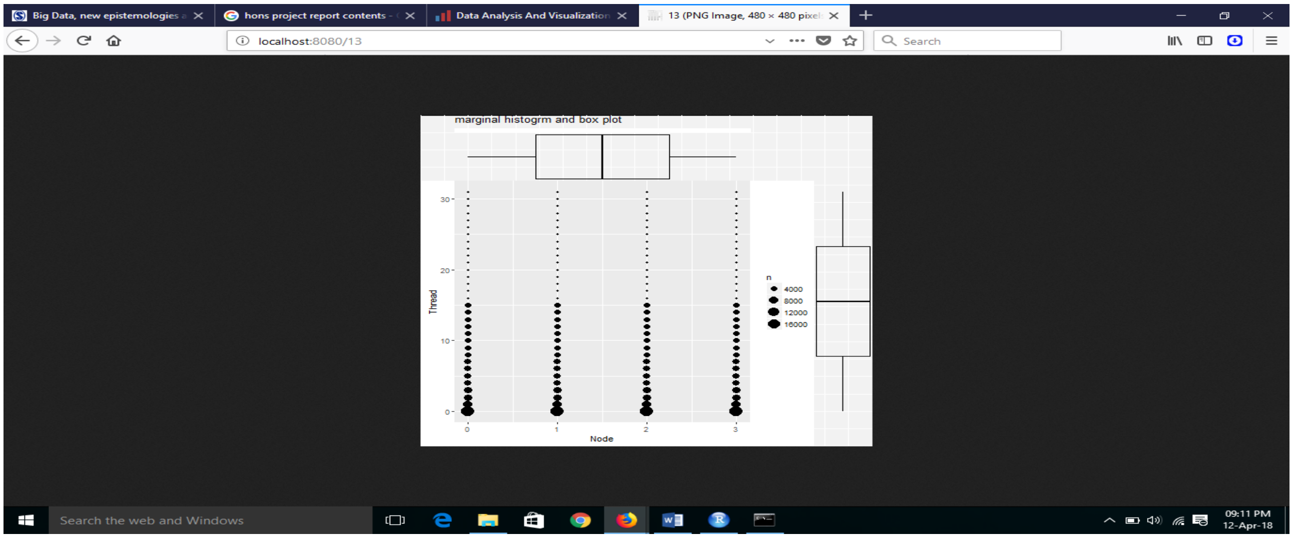
Big Data, new epistemologies
hons project report contents
Data Analysis And Visualization
13 (PNG Image, 480 × 480 pixels)
localhost:8080/13
Search
marginal histogrm and box plot
Thread
30
20
10
0
0
1
2
3
Node
n
4000
8000
12000
16000
Search the web and Windows
09:11 PM
12-Apr-18

Figure 2.9: MARGINALHISTOGRAM+BOX CHART for bigdata

2.4 Conclusion

From this study we have found that Visualizations can be static or dynamic. Interactive visualizations often lead to discovery and do a better job than static data tools. Do interactive visualization for more analyzing of data. Interactive brushing and linking between visualization approaches and networks or Web-based tools can facilitate the scientific process. Web-based visualization helps get dynamic data timely and keep visualizations up to date. More new methods and tools of Big Data visualization should be developed for different Big Data applications. Big Data analytics and visualization can be integrated tightly to work best for Big Data applications. More new graphs and More new visualizations could try in the future for better enhancement of visualization. A lot of technologies are developing for better visualization of bigdata and in future this can be developed by asking the user to enter data and ask like to which points data should be plotted.

Chapter 3

Literature Survey

- Early work in this area includes Turney [17] and Pang [18] who applied different methods for detecting the polarity of product reviews and movie reviews respectively. It is less clear how sentiment analysis techniques can be employed in the context of social website analysis where the language tends to be more freeform and informal.

- Siersdorfer analyzed more than 6 million comments collected from 67,000 YouTube videos to identify the relationship between comments, views, comment ratings and topic categories.

- Shaila S.G, Prasanna MSM, Kishore Mohit Have Done Classification Of YouTube Data Based On Sentimental Analysis. They Have Presented In International Journal Of Engineering Research In Computer Science And Engineering(ijercse)vol 5, Issue 6, June 2018.

- Pang, Lee And Vaithyanathan [8] Perform Sentiment Analysis On 2053 Movie Reviews Collected From The Internet Movie Database (Imdb).

Their Work Depicted That Standard Machine Learning Techniques Such As Nave Bayes Or Support Vector Machines (Svms) Outperform Manual Classification Techniques That Involve Human Intervention. However, The Accuracy Of Sentiment Classification Falls Short Of The Accuracy Of Standard Topic-based Text Categorization

- Mishne And Glace Focused On Comments In Webblogs . They Found That Comments Constitute A Substantial Part Of The Blogosphere, Accounting For Up To 30 percent Of The Volume Of Weblog Posts Themselves. Their Work Underlines The Importance Of User Comments As A Way How People Interact And Extend Primary Content. Even Though Web Blogs Are A Quite Different Domain, The Authors Identified Comment Types Which Are Very Similar To Those We Found In Our Study(e.g. Discussion).
- Chandramouli, R Have Done Sentiment Extraction From Text Can Be Used To Predict Mood, Which Can Be Used To Prevent Or Mitigate Security Threats.

A profound writing survey was directed in this investigation to find drifting themes, conceivable commitments of the related examinations and their future proposals to frame a reason for the exploration of this study. In this manner, to the best of creators' information, it very well may be concluded that:

- There isn't a multi-lingual system for Youtube comment sentiment analysis.

- lexicon based (dictionary based) assumption analysis is still most mainstream rather than machine learning, classification and grouping.

- Multicultural examination of online networking information on assumption investigation has not been done yet.

- Information accumulation is the slightest specified part in articles, while proposing a novel strategy for this issue can be extremely strong for the scholastics.

- Some of the dictionaries aimed to be used in this study are mentioned in some studies but have not been used all together yet (possibly because of huge work requirement).

- Big data studies are becoming very popular on sentiment analysis but have not been defined well yet.

Chapter 4

Proposed scheme and Architecture

4.1 Proposed Architecture

Architecture is getting comments from youtube API and directly importing to CSV file .Data cleaning and stop words are extracted from comments so that each word of sentiment calculated and printed their sentiment scores on the graph.

Diagram.png

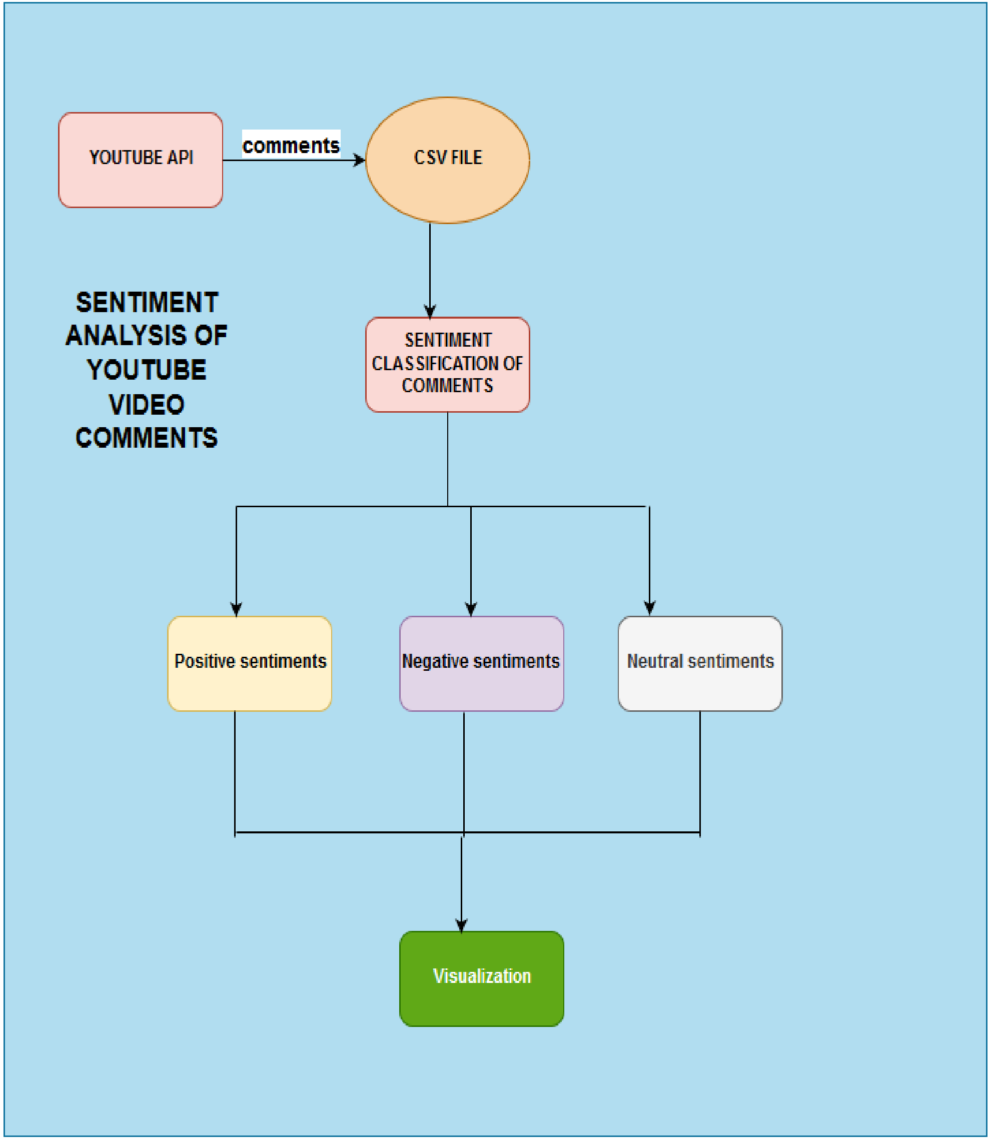

Figure 4.1: Proposed Architecture

4.2 Implementation

- getting Youtube comments from youtube Api with the help of api key.

- Deriving the sentiment of each comment: Sentiment for each comment was computed based on the sentiment score of the terms in the comment. The sentiment of a comment is equivalent to the sum of the sentiment scores for each term in the comment. The file nrc . containing a list of pre-computed sentiment scores was used to determine term score in each tweet.

- Deriving the sentiment of new terms: The sentiment for the terms that do not appear in the file nrc was computed based on the overall tweet sentiment deduced in step 2.

- Computing term frequency: Frequency of each term was calculated as [of occurrences of the term in all comments]/[of occurrences of all terms in all comments].

4.3 Network Analysis

Over a wide range of fields network analysis has become an increasingly popular tool for scholars to deal with the complexity of the interrelationships between actors of all sorts. The promise of network analysis is the placement of significance on the relationships between actors, rather than seeing actors as isolated entities.

There are a number of applications designed for network analysis and the creation of network graphs such as gephi and cytoscape. Though not specifically designed for it, R has developed into a powerful tool for network analysis. The strength of R in comparison to stand-alone network analysis software is three fold. In the first place, R enables reproducible research that is not possible with GUI applications. Secondly, the data analysis power of R provides robust tools for manipulating data to prepare it for network analysis. Finally, there is an ever growing range of packages designed to make R a complete network analysis tool. Significant network analysis packages for R include the statnet suite of packages and igraph.

4.3.1 Nodes and Edges

The two primary aspects of networks are a multitude of separate entities and the connections between them. The vocabulary can be a bit technical and even inconsistent between different disciplines, packages, and software. The entities are referred to as nodes or vertices of a graph, while the connections are edges or links. In this post I will mainly use the nomenclature of nodes and edges except when discussing packages that use different vocabulary.The network analysis packages need data to be in a particular form to create the special type of object used by each package. The object classes for network, igraph, and tidygraph are all based on adjacency matrices, also known as socio matrices.

4.3.2 edge and node lists

To create network objects from the database of letters received by Daniel van der Meulen in 1585 I will make both an edge list and a node list. This will necessitate the use of the dplyr package to manipulate the data frame of letters sent to Daniel and split it into two data frames or tibbles with the structure of edge and node lists.

4.3.3 Network Objects

The network object classes for network, igraph, and tidygraph are all closely related. It is possible to translate between a network object and an igraph object. However, it is best to keep the two packages and their objects separate. In fact, the capabilities of network and igraph overlap to such an extent that it is best practice to have only one of the packages loaded at a time. I will begin by going over the network package and then move to the igraph and tidygraph packages.

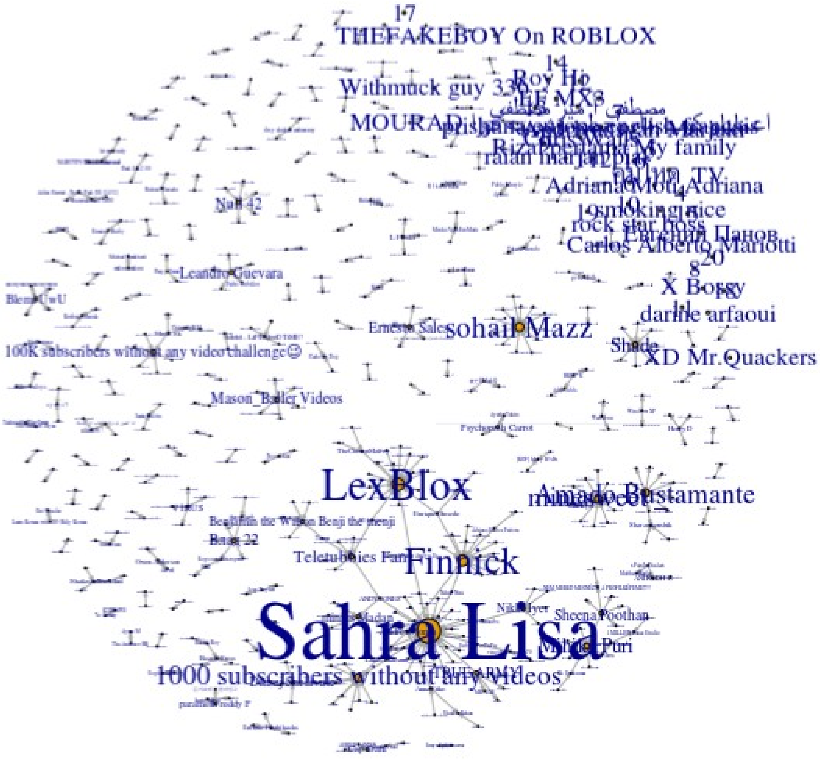

Figure 4.2: Network Object Diagram

Chapter 5

Results and Discussions

Lexicon oriented techniques for the detection of sentiment polarity are based on different lexicons, such as, WordNet and SentiWordNet (SWN),AFINN. For the detection of sentiments from user comments, an existing WordNet dictionary and a specific list are used. The specific list contains terms and phrases. These terms and phrases express user opinions.The SentiWordNet (SWN) is a document resource which contains a list of English terms which have been attributed a score of positivity and negativity[14].The SWN assigns polarity to each term and phrase of WordNet.For analyzing user sentiment in comments, SentiWordNet is used.

When a user enters a comment, the system detects the number of those words having sentiment. A sentiment score is calculated by the number of entries for every word present in the list. If there is a single entry of a word then that word is assigned a polarity with the highest score. If there are more than one entries of a term, the scores of each class are averaged for normalization. For example, there are four scores for a term in a sentence.

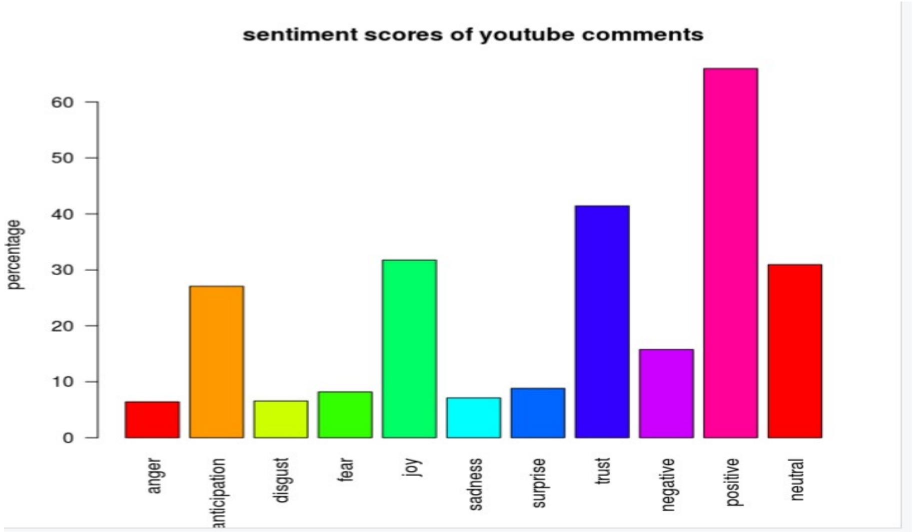

Figure 5.1: Barplot of sentimental analysis 1

To find the sentiment polarity of the entire comment, the average score of all terms in the comment is calculated and the term having highest frequency is assigned either positive or negative polarity. Table 1 shows sentiment scores for terms.

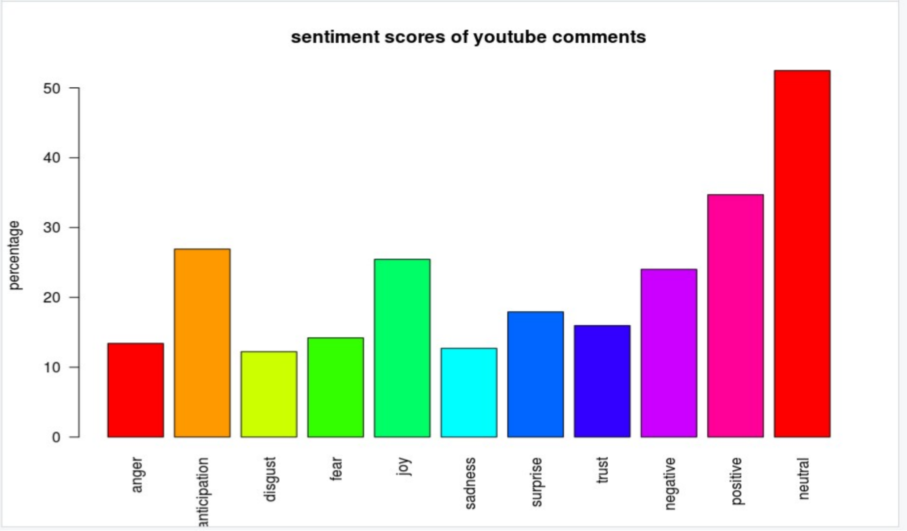

Figure 5.2: Barplot of sentimental analysis 2

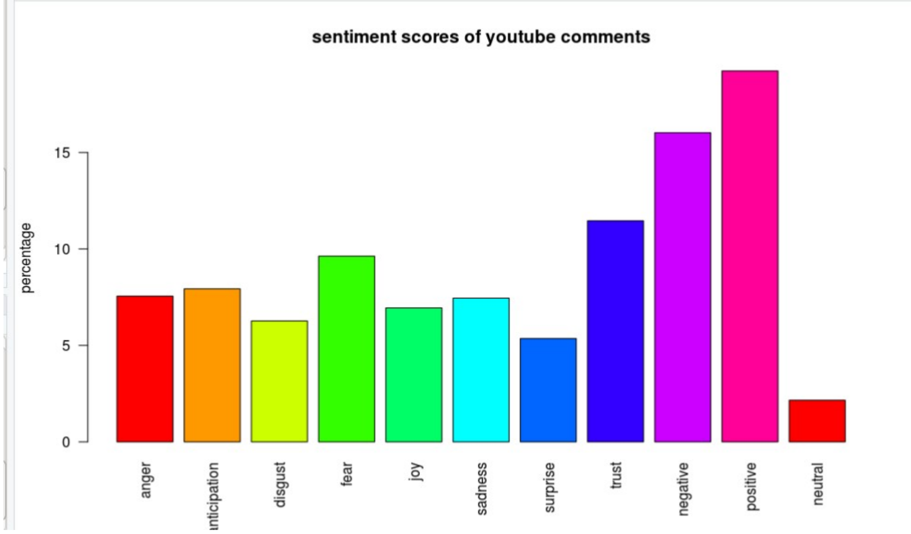

Figure 5.3: Barplot of sentimental analysis 3

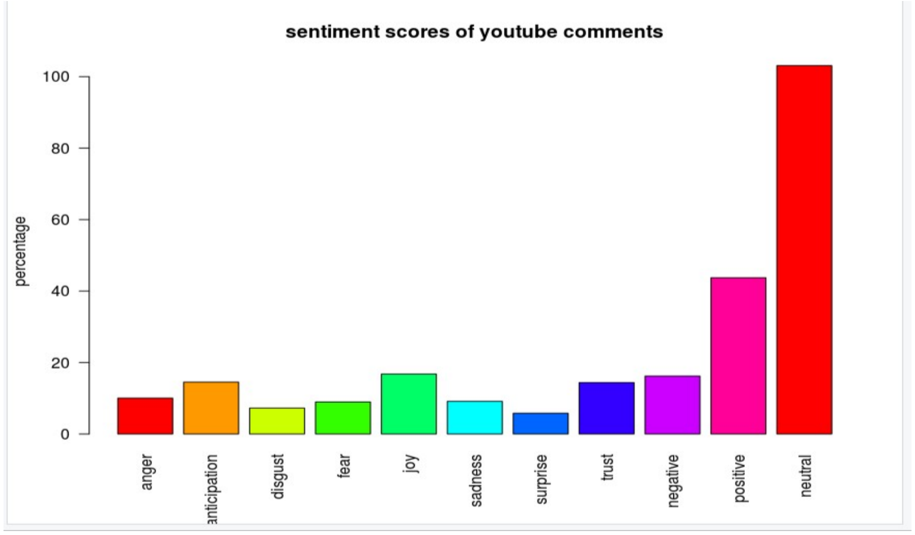

Figure 5.4: Barplot of sentimental analysis 4

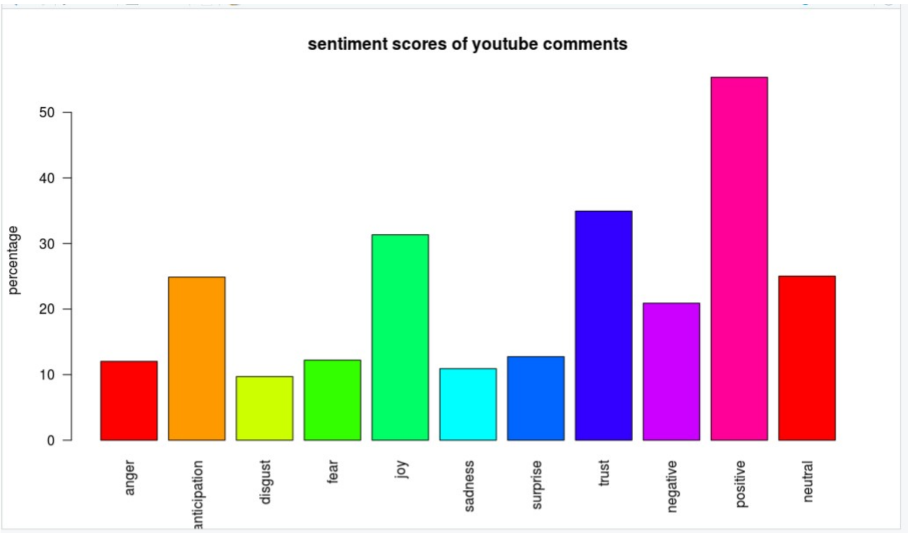

Figure 5.5: Barplot of sentimental analysis 5

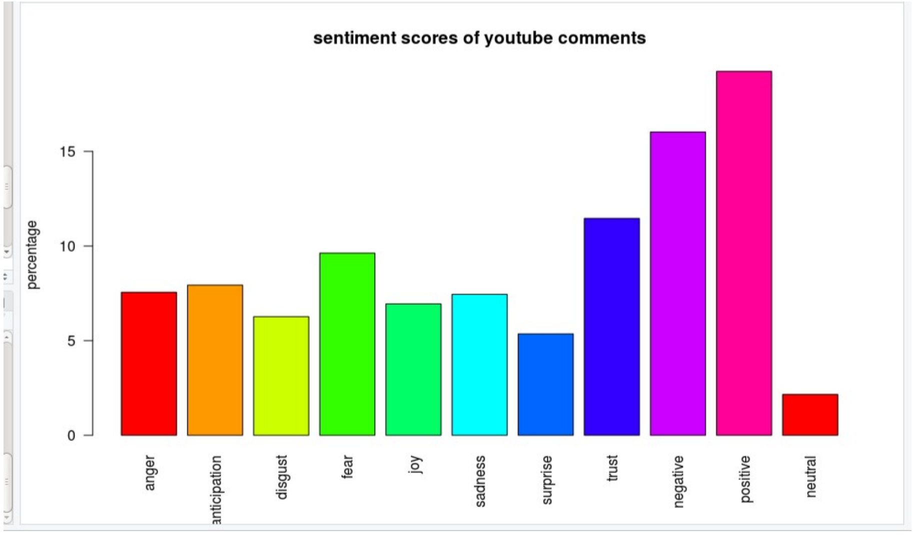

Figure 5.6: Barplot of sentimental analysis 6

Chapter 6

Conclusion

Classification of general events and detection of Sentiment Polarity of user comments in YouTube is a challenging task for researchers so far. A lot of work is done in this regard but still have a long way to go to overcome this problem. In this paper we have emphasized on following problems in order to find the polarity of comments given by the users of YOUTUBE.1) Current sentiment dictionaries having limitations.2) Informal language styles used by users, 3) Estimation of sentiments for community-created terms, 4) To assign proper labels to events, 5) Achieve satisfactory classification performance 6) Challenges involving social media sentiment analysis. Different techniques like User Sentiment Detection, Event Classification and Predicting YOUTUBE comments used for comments polarity are also discussed. Regarding future work, improving the social lexicon and to validate it statistically and proper event classification can help to increase the performance to predict rating of comments.

Bibliography

[1] Muhammad Zubair Asghar, Shakeel Ahmad, Afsana Marwat, and Fazal Masud Kundi. Sentiment analysis on youtube: a brief survey. *arXiv preprint arXiv:1511.09142*, 2015.

[2] Shangkun Deng, Takashi Mitsubuchi, Kei Shioda, Tatsuro Shimada, and Akito Sakurai. Combining technical analysis with sentiment analysis for stock price prediction. In *2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing*, pages 800–807. IEEE, 2011.

[3] Himaanshu Gauba, Pradeep Kumar, Partha Pratim Roy, Priyanka Singh, Debi Prosad Dogra, and Balasubramanian Raman. Prediction of advertisement preference by fusing eeg response and sentiment analysis. *Neural Networks*, 92:77–88, 2017.

[4] Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha. Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks*, pages 27–38. ACM, 2013.

[5] Haeng-Jin Jang, Jaemoon Sim, Yonnim Lee, and Ohbyung Kwon. Deep sentiment analysis: Mining the causality between personality-value-attitude for analyzing business ads in social media. *Expert Systems with applications*, 40(18):7492–7503, 2013.

[6] Amar Krishna, Joseph Zambreno, and Sandeep Krishnan. Polarity trend analysis of public sentiment on youtube. In *Proceedings of the 19th International Conference on Management of Data*, pages 125–128. Computer Society of India, 2013.

[7] Tao Li, Lei Lin, Minsoo Choi, Kaiming Fu, Siyuan Gong, and Jian Wang. Youtube av 50k: an annotated corpus for comments in autonomous vehicles. *arXiv preprint arXiv:1807.11227*, 2018.

[8] Peter Schultes, Verena Dorner, and Franz Lehner. Leave a comment! an in-depth analysis of user comments on youtube. *Wirtschaftsinformatik*, 42:659–673, 2013.

[9] Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. How useful are your comments?: analyzing and predicting youtube comments and comment ratings. In *Proceedings of the 19th international conference on World wide web*, pages 891–900. ACM, 2010.

[10] Mike Thelwall. The heart and soul of the web? sentiment strength detection in the social web with sentistrength. In *Cyberemotions*, pages 119–134. Springer, 2017.

[11] Mike Thelwall and Kevan Buckley. Topic-based sentiment analysis for the social web: The role of mood and issue-related words. *Journal of the*

American Society for Information Science and Technology, 64(8):1608–1617, 2013.

[12] Mike Thelwall, Pardeep Sud, and Farida Vis. Commenting on youtube videos: From guatemalan rock to el big bang. *Journal of the American Society for Information Science and Technology*, 63(3):616–629, 2012.

[13] Olga Uryupina, Barbara Plank, Aliaksei Severyn, Agata Rotondi, and Alessandro Moschitti. Sentube: A corpus for sentiment analysis on youtube social media. In *LREC*, pages 4244–4249, 2014.