ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

Sentiment Analysis On Unstructured Data

[1] Apurva Joshi, [2] Kalyani Birgade, [3] Pallavi Petkar [4] Mrunali Sathone
[1][2][3][4] Assistant Professor
Department of Information Technology,
Yeshwantrao Chavan College of Engineering Nagpur (MS), India

***Abstract* - An enormous growth of the WWW has been instrumental in spreading social networks. Due to many-fold increase in internet users taking to online reviews and opinions, the communication, sharing and collaboration through social networks have gained importance. The rapid growth in web-based activities has led to generation of huge amount of unstructured data which accounts for over 80% of the information. Exploiting big data alternatives in storing, processing, archiving and analyzing this data becomes increasingly necessary. Unstructured data refers to information that either does not have predefined data model or is not organized in a predefined manner. Unstructured data is being constantly generated via call center logs, emails, documents on the web, blogs, tweets, customer comments, customer reviews and so on. While the amount of data is increasing rapidly, the ability to summarize, understand and make sense of such data for making better decision remain challenging. So thus there is a need of sentiment analysis on unstructured data. In this paper we are describing what is sentiment analysis and methodology of analyzing on unstructured data .We have done analysis on various data sets from twitter , blogs and movielen.com site using r statistical language and output are visualized in the form of word cloud and histogram.We have created GUI's for analysis of this datasets by which users can easily analyze the data.**

***Index Terms:*— R, sentiment analysis,word cloud;**

I. INTRODUCTION

Social network have changed the way in which people communicate. The use of social networks allows for the sharing of content between people in an easier ,faster way. People express themselves through text ,music and videos on various sites such as Twitter, Facebook, Blogs etc .Information available from social network is used for analysis of peoples opinion for example looking at a response of people for particular product we can decide whether to buy it or not. But as this data is growing fastly and is overloaded, it is tedious task to analyze manually .

Unstructured data is a data which does not reside in column row format. Unstructured textual data is being constantly generated via call center logs, emails, documents on the web, blogs, tweets, customer comments, customer reviews, and so on. While the amount of textual data is increasing rapidly, ability to summarize, understand, and make sense of such data for making better decisions remain challenging. This paper takes a quick look at how to organize and analyze textual data from a large collection of documents and for using such information to improve performance.

Sentiment analysis is all about opinion mining implies extracting opinions ,emotions and sentiment from data. It is used to track attitudes and feelings on the web specially for measuring performance of product, services ,and brands. Sentiment analysis is the process of identifying whether a piece of writing is positive, negative , neutral .It is known as opinion mining because it derives opinion or attitude of a speaker. Sentiment analysis is a new area which deals with extracting user opinion automatically.

Consider an example of positive sentiment is "Asus Zenfone2 is a good model." Alternatively a negative sentiment is "Asus Zenfone2 is not working properly." So accordingly we can analyze the views of people. So in this paper we are gathering the data of particular entity, person from different data sources i.e.twitter,blogs,movie data set and analyzing it and producing a output which shows negative and positive sentiment of that entity, person or a word cloud which describes the data.

II. LITERATURE SURVEY

There are actually many sources of unstructured data, which accounts for much of the data growth that we've seen in the enterprise, academics and social media. Experts estimate that 80 to 90 percent of the data in any organization is unstructured. And the amount of unstructured data in enterprises is growing significantly.

1. ***Sentiment Analysis and Feedback Evaluation***
Feedback Evaluation is a necessary part of any

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

institute to maintain and monitor the academic quality of the system. Traditionally, a questionnaire based system is used to evaluate the performance of teachers of an institute. Here, they propose an automatic evaluation system based on sentiment analysis, which shall be more versatile and meaningful than existing system. In their proposed system, feedback is collected in the form of running text and sentiment analysis is performed to identify important aspects along with the orientations using supervised and unsupervised machine learning Above paper is analyzing the feedback and according a teacher graded in school.

2. ***Sentiment Analysis of Movie Reviews and Blog Posts*** This paper presents their experimental work on performance evaluation of the SentiWordNet approach for document-level sentiment classification of Movie reviews and Blog posts. They have implemented SentiWordNet approach with different variations of linguistic features,scoring schemes and aggregation thresholds. They have used two pre-existing large datasets of Movie Reviews and two Blog post datasets on revolutionary changes in Libya andTunisia. We have computed sentiment polarity and also its strength for both movie reviews and blog posts. The paper also presents an evaluative account of performance of the SentiWordNet approach with two popular machine learning approaches: Naïve Bayes and SVM for sentiment classification.

3. ***Extracting New Product Ideas from Consumer Blogs*** This paper introduces a web mining approach for automatically identifying and extracting new product ideas from internet blogs. There are a large amount of web logs for nearly each topic, where consumers present their needs for new products. These new product ideas are valuable to producers, researchers and developers because they can lead to a new product development process, a well-known task in marketing. However, the current approach towards extraction of these new product ideas involves employed analysts to process these blogs manually and extract the new product ideas. Here, presented is an automated approach towards the extraction of new product ideas through knowledge extraction and using scripting languages, which not only eliminates human error, but also speeds up the entire process of new product idea extraction.

4. ***An Intelligent Framework for Text-to-Emotion Analyzer*** This paper proposes an intelligent framework to detect the emotion of a text. Automatic derivation of the emotion from text is a challenge as it minimizes the misunderstanding by conveying the internal state of the users. They divide the framework into two modules, namely Training Module and Emotion Extraction Module. They utilize the concept of Exploratory Data Warehouse (DW) technology to train system. Therefore, DW relies not only on internal data but also on external (Web) data. The DW is used by the Emotion Extraction Module to detect the emotion of a given text.

III. METHODOLOGY

We have extracted data from three different sources i.e.twitter, blogs, movielen website.So first dataset is twitter.

1. ***Twitter Data Set:*** Flowchart of analyzing tweets of Twitter:

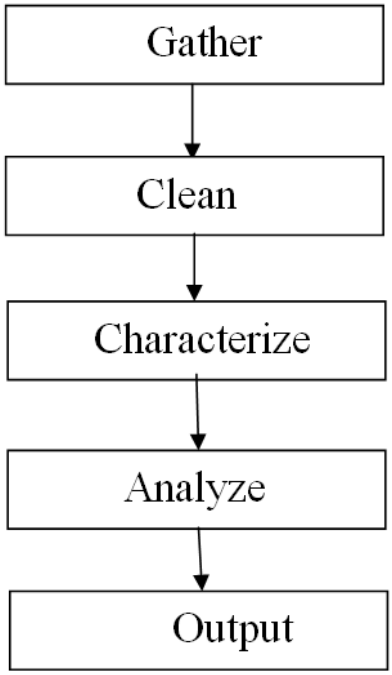

Figure 1: Flowchart of analyzing tweets of Twitter

1: Gather Tweets

Collecting tweets on a particular entity or person to analyze .

2: Clean Tweets

White spaces, special symbols, numbers are removed.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

3: Characterize Tweets

Characterize the sentiments in positive, negative and neutral

4: Analyze Tweets

Calculate the score of positive ,negative and neutral sentiment

5: Output

In any desirable format like histogram, pie diagram, word cloud format, frequency distribution, mean ,median etc.

Firstly to access tweets from twitter we have to do handshake authorization with twitter using our own twitter account using following code :

consumer_key <-'xyz'
consumer_secret <-'abc'
access_token <-'asd'
access_secret <-'fdg'
setup_twitter_oauth (consumer_key, consumer_secret,access_token,access_secret) the consumer key ,consumer secret ,access token ,access secret are private account keys.

Following are the R commands to execute above process:

1.Gather Tweets:

To gather tweets from twitter first we install following packages
install.packages(twitteR)
install.packages(RCurl)
we can search required tweets using following command
tweets <- searchTwitter("corruption",n=10,lang="en")

2.Clean Tweets:

To clean tweets Plyr and StringR package are required.
All the spaces,punctuation marks ,special symbols are removed.
install.package(plyr)
install.package(stringR)
then code for removing unwanted things is

```
sentence <- gsub('[[:punct:]]', "", sentence)
sentence <- gsub('[[:cntrl:]]', "", sentence)
sentence <- gsub('\\d+', "", sentence)
sentence <- tolower(sentence)
word.list <- str_split(sentence, '\\s+')
words <- unlist(word.list)
```

3.Characterize Tweets:

To characterize positive and negative words are separated and stored and matched with dictionary of positive and negative words.

```
pos.matches <- match(words, pos.words)
neg.matches <- match(words, neg.words)
pos.matches <- !is.na(pos.matches)
neg.matches <- !is.na(neg.matches)
```

4.Analyze Tweets

To analyze tweets function is used
analysis <score.sentiment(tweets_txt,pos.words,neg.words)

5.Output

To display the results:
hist(analysis$score)

2.Blogs Dataset

Flowchart of analyzing blogs:

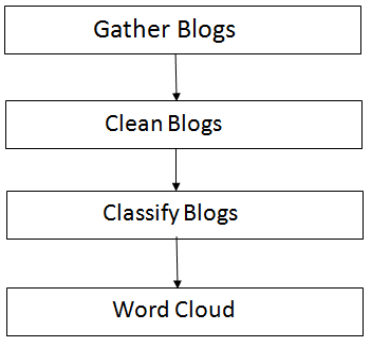

Figure 2: Flowchart of analyzing blogs

1.Gather Blogs:

Blogs are extracted from the site blogger.com in the .csv format.

2.Clean Blogs:

White spaces, special symbols, numbers ,unwanted words are removed.

3.Word Cloud:

Output of the analysis is visualized in form of word cloud.
Following are the R commands to execute above process:

- This are packages required to extract, clean, analyze and create a word cloud:

require("tm")
require("SnowballC")

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

```
require("wordcloud")
require("RColorBrewer")
```

- To clean the data following commands are required:

```
jeopCorpus<-Corpus(VectorSource(jeopQ$genres))
tdm = TermDocumentMatrix(jeopCorpus,
control = list(remove
Punctuation = TRUE,
stopwords = c("the","this", stopwords("english")),
removeNumbers = TRUE,
tolower = TRUE))
```

- Word cloud is created by :

```
wordcloud(jeopCorpus, min.freq = 1,
max.words=200,random.order=FALSE,  rot.per=0.35,
colors=brewer.pal(8, "Dark2"))
```

3.Movie Dataset

Flowchart of analyzing movie dataset:

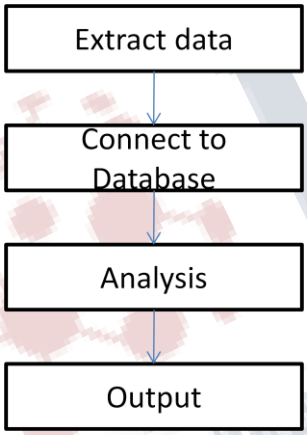

Figure 3: Flowchart of analyzing movie dataset

1.Extract data

Extract the data from movielen.com website in .csv format.The dataset contains three .csv files i.e. movies.csv which contains name and the respective genres of movie.

2.Connect to Database

Connection to sqlite database.

3.Analysis

Analysis is done according to the classification of genres of movies or according to year in which this releases,etc.

4.Output

Output is visualized in the form of wordcloud and histogram. Commands require by this analysis is same as blog analysis. Also the GUI is prepared for twitter, blog and movie analysis through which user can easily analyze the data for their use.This is created in R using the package called "shiny".In shiny,two files need to be created ui.r and server.r. Both files are connected ui file shows the display and server file computes the analysis.

IV. RESULTS

1.Twitter Dataset:

The output for twitter analysis is

```
> analysis
   score
1      0
2     -2
3     -2
4     -1
5      0
6      1
7     -1
8     -1
9     -2
10    -1
```

Figure 4: In tabular format

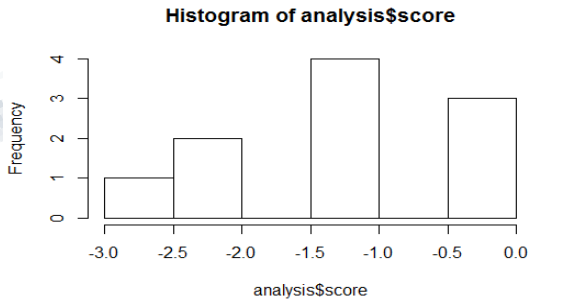

Figure 5: Histogam

Above is the output for the code in which the tweets are scored and according to that histogram is drawn.

2.Blogs Dataset

The outputs for blog analysis are:

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

Figure 6:Word Cloud

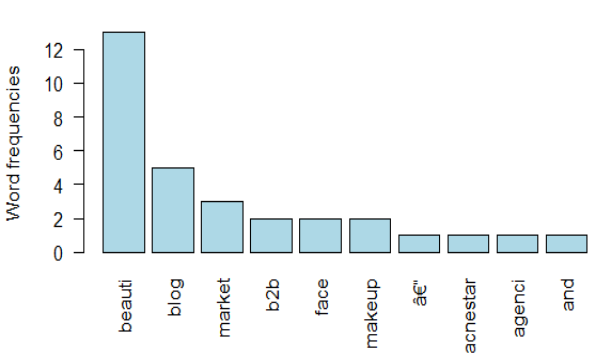

Figure 7:Histogram for blogs

word	freq
beauti	13
blog	5
market	3
b2b	2
face	2
makeup	2
â€"	1
acnestar	1
agenci	1
and	1
archive	1
attitud	1
bailylamb	1
blogger	1
bore	1
bright	1
costeffect	1
coutur	1
creat	1
denim	1
diari	1
dreamer	1
dress	1
email	1
fan	1
find	1
gee	1
glow	1
great	1
guid	1
gym	1
help	1
holiday	1
how	1
hustl	1
internet	1

Figure 8: Freqeuncies in table format

3.Movielen Dataset

The outputs for movie dataset analysis are:

Figure 9: Word cloud for genres of movie

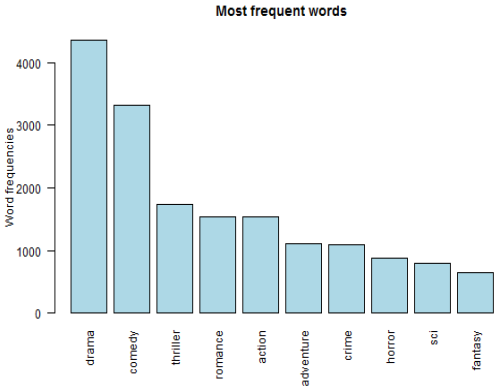

Figure 10:Histogram for movie dataset

Frequencies for the same:

word	freq
drama	4365.00
comedy	3315.00
thriller	1729.00
romance	1545.00
action	1545.00
adventure	1117.00
crime	1100.00
horror	877.00
sci	792.00
fantasy	654.00

Figure 11: Frequencies for the same

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

4.GUI's

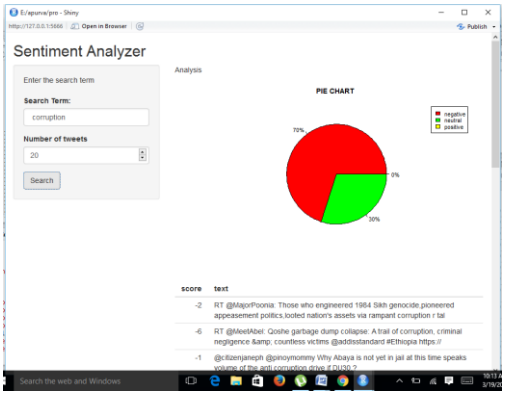

Figure 12: GUI for twitter analysis

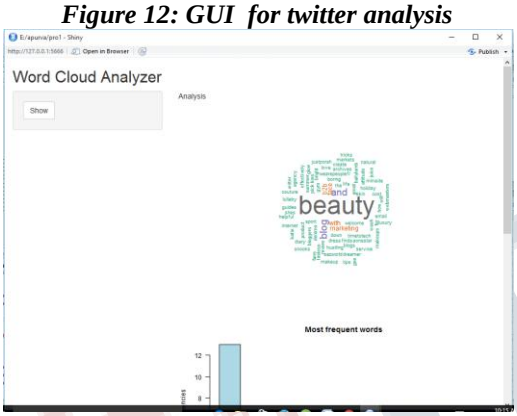

Figure 13: GUI for blog analysis

Figure 14: GUI for movie dataset analysis

V ACKNOWLEDGMENT

For the successful completion of this project we would like to thank Prof. Rupa Fadnavis mam .Thanks are also extended to Prof. K.HAJARI sir for his support during the project.

VI CONCLUSION

In this era of huge burst of data, it is essential to make use of that data in efficient manner. This can be done by analyzing the data coming from different sources, so as to take good decisions in different feilds. In this paper ,we have discussed about what is sentiment ,what is unstructured data and the methodologies of analysis on unstructured data using R language.We have data from three different sources i.e twitter ,blogs, movie dataset. We have created the GUI's where this datasets can be analyzed and the desired outputs can be seen. This will help the users to use the big data in efficient manner.

REFERENCES

Jalaj S. Modha, Prof. and Head Gayatri S. Pandi, Sandip J. Modha, ―Automatic Sentiment Analysis for Unstructured Data‖, International Journal of Advanced Research in Computer Science and software Engineering, Volume 3, Issue 12, December 2013 pp no (91-97)

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proc. ACL 2011 Workshop on Languages in Social Media. pp. 30–38 (2011)
Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In:Proceedings of COLING. pp. 36–44 (2010)

Davidov, D.; Tsur, O.; and Rappoport, A. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In Proceedings of Coling.

V. K. Singh, M. Mukherjee and G. K. Mehta, "Combining Collaborative Filtering and Sentiment Analysis for Improved Movie Recommendations", In C. Sombattheera et. al. (Eds.):

Multi-disciplinary Trends in Artificial Intelligence, LNAI 7080,Springer-Verlag, Berlin-Heidelberg, pp. 38-50, 2011.

V. K. Singh, M. Mukherjee and G. K. Mehta, "Sentiment andMood Analysis of Weblogs using POS Tagging based Approach",In S. Aluru et al. (Eds.): IC3 2011, CCIS 168, pp. 313-324,Springer-Verlag, Berlin Heidelberg, 2011.

N. Godbole, M. Srinivasaiah, S. Skiena. Large-Scale Sentiment Analysis for News and Blogs. In proceedings of ICWSM 2007, pages 219.