IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.7, Special Issue.4, Feb. 2019 E-ISSN: 2347-2693

Social Media Mining : Retrieving , Preprocessing Storing and Analyzing Bone Cancer Related Tweets Using R

S. Mahalakshmi

Department of Computer Science, R.A College for Women, Thiruvarur, India

Corresponding Author : laxmigaya22@gmail.com

Available online at: www.ijcseonline.org

***Abstract*—**Social media provides easily an accessible platform for users to share information. Mining social media has its potential to extract actionable patterns that can be beneficial for business, users,and consumers. Social media data are vast, noisy, unstructured, and dynamic in nature, and thus novel challenges arise. This paper deals with social media mining in which we retrieved tweets ,preprocessed and store it in a csv file in order to compare with ontology related to cancer which is created using protégé. Also analysis made on preprocessed cancer related tweets using R.

***Keywords*—**Social media, Mining, preprocess, csvfile, Ontology, Tweets, R

I. INTRODUCTION

Social media plays a crucial role in every aspect and in all corners of the world. Many social medias like facebook, twitter, myspace etc., are extensively used now a days. We used tweets from twitter to carry our research work.

The use of social media for health monitoring and surveillance has indeed many drawbacks and difficulties, particularly if done automatically. For example, traditional NLP methods that are used on longer texts have proven to be inadequate when applied to short texts, such as those found in Twitter. Something seemingly simple, such as searching and collecting relevant postings has also proven to be quite challenging, given the amount of data and semantic heterogeneity (how people refer to the topic of interest in colloquial terms) inherent to the media.

II. RELATED WORK

The pervasive use of social media sites such as Facebook, Instagram, LinkedIn, and Twitter have been producing great amounts of a new form of data, simply known as social media data. It is mostly user-generated, informal, incomplete, and multi-media, and is often accompanied with information about time and location.

Data preprocessing is done to eliminate the incomplete, noisy and inconsistent data. Data must be preprocessed in order to perform any data mining functionality.[2].

Microblogging platforms are used by different people to express their opinion about different topics, thus it is a valuable source of people's opinions. Twitter contains an enormous number of text posts and it grows every day. The collected corpus can be arbitrarily large. Twitter's audience varies from regular users to celebrities, company representatives, politicians, and even country presidents. Therefore, it is possible to collect text posts of users from different social and interests groups. Twitter's audience is represented by users from many countries.[3].

Social networks have seen an unprecedented growth in terms of users worldwide (e.g., as of 11th July 2014, Twitter has over 645 million users and grows by an estimated 135,000 users every day, generating 9,100 tweets per second). The social networks form a platform for people to share, discuss, and update their views and opinions, and many share their health-related information both in generic social media (such as Twitter, Facebook or Instagram) and in health-related social networks (forums focusing specifically on health issues, such as DailyStrength or MedHelp). Advances in automated data processing, machine learning and natural language processing present and attributes of the concept known as slots and constrains on these slots .

Ontologies play an important role in semantic description for common understanding and classification of the documents in the knowledge domain . They use a single concept for reducing ambiguous concepts or terminology and support the exchange of information retrieval; they are also critical to the development of the knowledge based systems [4].

Researchers have shown that finding information is an important use for status updates. For example, Lampe et al. [5] found university students used Facebook "to get useful

information." Java et al. [6] identified "information seekers" as a primary category of Twitter users. Zhao and u"gathering useful information for one"s profession or other personal interests," and "seeking for help and opinions," and Naaman et al. [8] found that questions to followers made up about 5% of posts that they manually coded. Honeycutt and Herring [9] similarly found that tweets directed at specific Twitter users were sometimes meant to "solicit information." Morris et al. [10] explored the use of status messages to find information by asking questions.

III. TWITTER MINING METHODOLOGY

Twitter is a social networking site that allows users to send and read short messages of a maximum of 140 characters. Twitter was created in March 2006 and was officially launched in July 2006. The growth of Twitter has been phenomenal, currently having reached over 200 million users and handling over 200 million tweets per day.

Users on Twitter are identified by a user name, and this user name is proceeded by the "@" symbol. When a user identifies another user in their tweet by their user name, it will be visible to the public, and the user that is referenced will be notified by Twitter that they have been "mentioned." If a user sees a tweet that is interesting and wants to pass the information along, they can "retweet" the post, which is similar to forwarding an email message to a new set of users, in this case their followers. Retweets will generally be identified with an "RT" that is embedded in the message. Lastly, messages can be grouped by topic or type by the use of hashtags (#). A hashtag preceding the topic will allow Twitter users to find tweets related to a particular topic when performing a search.

In our research we are retrieving and analyzing tweets related to different types of cancer .In this paper we presented and analysed bone cancer related tweets as follows.

A. Creating Twitter App

In order to retrieve tweets from Twitter API we must get OAuth keys by doing the following steps .Once We've done this, make a note of OAuth settings. we will need these long horrible strings of characters for our Twitter app. Theyare:

- Consumer Key
- Consumer Secret
- OAuth Access Token
- OAuth Access Token Secret

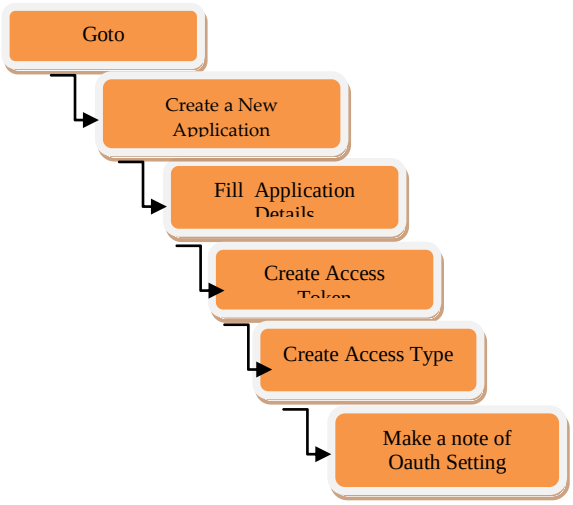

Fig 1.Steps to create Twitter App

B.About Rstudio

RStudio is a free and open sourceintegrated development environment (IDE) for R, a programming language for statistical computing and graphics.

Sample Tweets retrieved related to bone cancer in Rstudio

```
>library(twitteR)
>Setup_twitter_oauth('ConsumerKey','Consumer Secret','AccessToken',"Access Secret')
>tweetmaha<-searchTwitter('bone cancer',n=100)
>tweetmaha[1:5]
```

```
[[1]]
[1] "MangaiH: September:#ChildhoodCancerAwarenessmonth;different v adults. childhood cancer mostly effect white blood cells brain,bone,lymphatic system"
[[2]]
[1] "KoolnewsBracken: @AliGorman6abc, My cousin passed away years ago. What causes bone cancer?"
[[3]]
[1] "amyy_claire: My poor baby Allis has bone cancer, but she's still pretty happy. http://t.co/SQtQfq4ZT0"
[[4]]
[1] "RenataBeamanPT: #exercise #cancer Did you know exercise can reduce cancer fatigue, improve bone density, and improve skin health? https://t.co/PEWVN2O89I"
[[5]]
[1] "insurancehnews: New potential therapeutic strategy against a very aggressive infant bone cancer, metastatic Ewing sarcoma http://t.co/YrwknpQyD0"
```


C.Transforming Text

The tweets are first converted to a data frame and then to a corpus.

```
>df<- do.call("rbind", lapply(rdmTweets, as.data.frame))
> dim(df)
> library(tm)
> # build a corpus, which is a collection of text documents
> # VectorSource specifies that the source is character vectors.
>myCorpus<- Corpus(VectorSource(df$text))
```

After that, the corpus needs a couple of transformations, including changing letters to lower case, removing punctuations/numbers and removing stop words.

D.Preprocessing

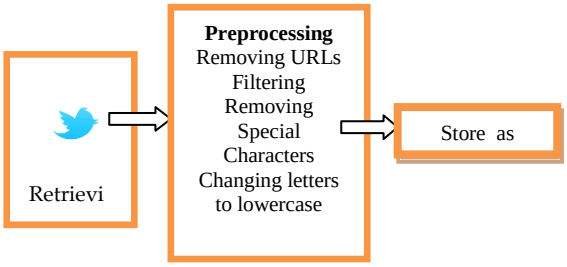

Fig2.Preprocessing

Data preprocessing is done to eliminate the incomplete, noisy and inconsistent data. Data must be preprocessed in order to perform any data mining functionality. Data Preprocessing involves the following tasks

- Removing URLs
- Filtering
- Removing Special Characters @,#etc
- Removal of unnecessary English words

Like his ,her,theetc

Coding in R for Preprocessing

```
>myCorpus<- tm_map(myCorpus, tolower)
>myCorpus<- tm_map(myCorpus, removePunctuation)
>myCorpus<- tm_map(myCorpus, removeNumbers)
>removeURL<-function(x)gsub("http[[:alnum:]]*","",x)
>myCorpus<-tm_map(myCorpus,removeURL)
>myCorpus<- tm_map(myCorpus, removeWords, stopwords('english'))
>myCorpus<- tm_map(myCorpus, stripWhitespace)
>corpus<- tm_map(myCorpus, stemDocument)
>corpusT<- tm_map(corpus, PlainTextDocument)
>dtm<-
TermDocumentMatrix(corpusT,control=list(minWordLength
=1))
>dtm
<<TermDocumentMatrix (terms: 134, documents: 50)>>
Non-/sparse entries: 558/6142
Sparsity           : 92%
Maximal term length: 17
Weighting          : term frequency (tf)
> frequent_ge_2 <- findFreqTerms(dtm, lowfreq = 2)
> frequent_ge_2
[1] "amp"         "andrew"      "basketball"  "battling"    "blair"
"bone"
 [7] "butler"      "can"         "cancer"      "capsules"    "care"
"cartilage"
[13] "cells"       "center"      "childhood"   "danafarber"
"depleted"    "dies"
[19] "educational" "event"       "fallujah"    "families"
"former"      "friend"
[25] "gistsoft"    "giving"      "health"      "help"        "howd"
"immix"
[31] "indystar"    "iraq"        "kids"        "legs"        "mines"
"mrmalky"
[37] "natural"     "old"         "patients"    "people"      "please"
"recurrence"
[43] "retweet"     "sarcoma"     "shark"       "shells"      "smith"
"son"
[49] "symposium"   "thx"         "tissue"      "tony"
"uranium"     "warcriminal"
[55] "years"       "young"
```

E.SystemArchitecture.

In this paper data's are retrieved from Twitter API and they are preprocessed in Rstudio and the preprocessed tweets related to bone cancer are stored as a csvfile.then these key words are compared with the ontology created related to bone cancer. Our proposed system first identifies the concepts and their relationship for the Cancer ontology. The identified concepts and their relationship are built into domain ontology with sufficient annotation properties and instances. After that we are going to classify and predict the tweets.

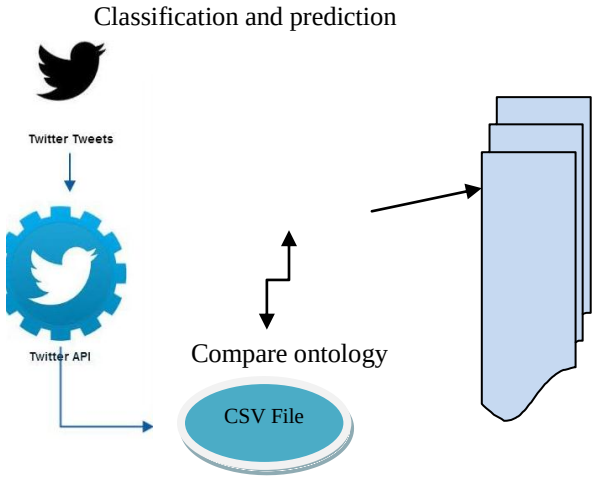

Fig3.System Architecture

IV. RESULTS AND DISCUSSION

We retrieved 100 tweets related to bone cancer and the key words are pick out from tweets after preprocessing and in Rstudio we analyse the datas and draw the graph as follows:

Analysing Term frequency in Rstudio using ggplot2 package

```
>term.freq<-rowSums(as.matrix(dtm))
>term.freq<-subset(term.freq,term.freq>=2)
>library(ggplot2)
>df<-data.frame(term=names(term.freq),freq=term.freq)
ggplot(df,aes(x=term,y=freq))+geom_bar(stat="identity")+xlab("terms")+ylab("count")+coord_flip()
```

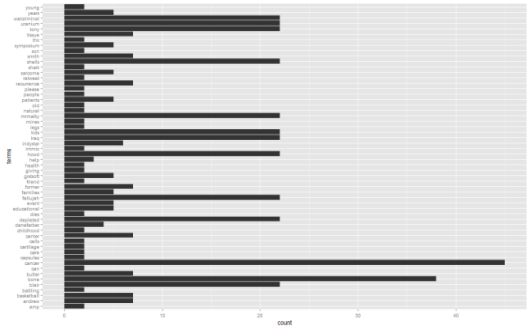

Fig 4.Graph using R count Vs Terms

Steps to get wordcloud inRstudio using Wordcloud package

Coding

```
>library(wordcloud)
> m<-as.matrix(dtm)
>word.freq<-sort(rowSums(m),decreasing=T)
>wordcloud(words=names(word.freq),freq=word.freq,min.freq=3,random.order=F)
```

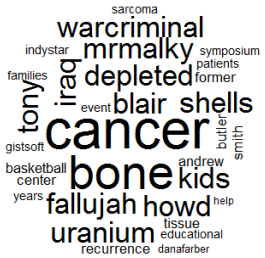

Fig 5.Word cloud

V. CONCLUSION AND FUTURE SCOPE

This paper deals with twitter mining in which we retrieved bone cancer related tweets ,preprocessed twitter data's using Rstudio .Then from these preprocessed tweets , word frequency was find out and those words are stored in a CSV file. In future we planned to compare the words in csv file with the bone cancer related words in ontology we created in protégé. .Then we are going to classify & predict the tweets according to the comparision results and also we decided to classify tweets for large volumeness of data with different types of cancers.

REFERENCES

[1] Inna Novalija, Miha Papler, Dunja Mladenić ,"Towards Social Media Mining: Twitterobservatory,Artificial Intelligence Laboratory ,Jožef Stefan Institute. Jamova 39, 1000 Ljubljana, Slovenia

[2] S. Tamilarasan, P.K. Sharma, "A Survey on Dynamic Resource Allocation in MIMO Heterogeneous Cognitive Radio Networks based on Priority Scheduling", International Journal of Computer Sciences and Engineering, Vol.5, No.1,pp.53-59, 2017.

[3] I.Hemalatha1 Dr. G. P Saradhi Varma2 Dr. A.Govardhan," Preprocessing the Informal Text for efficient Sentiment Analysis", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) ,Volume 1, Issue 2, July – August 2012 .

[4] Akshi Kumar and Teeja Mary Sebastian," Sentiment Analysis on Twitter", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012

[5] Pasapitch Chujai, Nittaya Kerdprasop, and Kittisak Kerdprasop, "On Transforming the ER Model to Ontology Using Protégé OWL Tool", International Journal of Computer Theory and Engineering, Vol. 6, No. 6, December 2014.

[6] Lampe, C., Ellison, N. and Steinfield, C. (2008). Changes in use and perception of Facebook. CSCW 2008, 721-730.

[7] Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we Twitter: Understanding microblogging usage and communities. WebKDD/SNA-KDD 2007, 56-65.

[8] Zhao, D. and Rosson, M.B. (2009). How and why people Twitter: The role that microblogging plays in informal communication at work. Group 2009, 243-252.

[9] Naaman, M., Boase, J., & Lai, C. H. (2010). Is it really about me? Message content in social awareness streams. CSCW 2010, 189-192.

[10] Honeycutt, C. and Herring, S. (2009). Beyond microblogging: Conversation and collaboration via Twitter. HICSS 2009.

[11] Morris. M.R., Teevan, J., and Panovich, K. (2010). What do people ask their social networks, and why? A survey study of status message Q&A behavior. CHI 2010, 1739-1748.

