ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

Analysing the Data from Twitter using R

K. Sailaja Kumar[1], D. Evangelin Geetha[1], T. V. Sai Manoj[2]

Department of Computer Applications, MSRIT, Bangalore, India [1]

Department of Computer Science and Engineering, MSRIT, Bangalore, India[2]

Abstract: Online Social Networks (OSNs) are the powerful medium for communication among the individuals to share their views on disastrous events happening in and around using the opportunities offered by the internet. This paper aims to analyze the meaningful real-time data about the disastrous events obtained from the most popular microblogging OSN 'Twitter'. Tweets related to the target event are gathered based on the search query, extracted the keywords from the tweets and then analyzed the significance of those keywords in the events happened during and after the disaster using text mining. The data visualization analytics supported by the statistical software tool 'R' is used to explain the discovered phenomena. Tweets are collected on 'Jammu and Kashmir Floods' using Twitter API based on various search queries and around 1570 tweet messages were examined. The obtained corpus is then processed using text mining functions provided in 'R'. A term document matrix is constructed to know the most frequent words, the distribution of the word frequencies and the association between them. The barplot is plotted to visualize the frequent words. Further the most popular keywords in the tweets and terms contained in the keywords are visualized by constructing a wordcloud from the term document matrix.

Keywords: Disastrous Events, Online Social Networks (OSNs), R, Term Document Matrix, Word cloud

I. INTRODUCTION

OSNs are the prevailing means of communication to share valuable knowledgeamong millions users. The most popular OSN Twitter is used as a source of information for research related to OSN domains. The studies have shown that Twitter data plays a major role on disaster management, mostly on disaster respond than disaster relief or post event [1]. Twitter research evolves tools and methods for capturing data (tweets) during and after the disastrous events [2]. This will be useful in coordinating resources and efforts and also in preparing and planning for disaster relief.

Twitter data can be obtained through the Application Programming Interface (API) provided by the Twitter. To analyse the Twitter data we needdata analysis tools such as R. 'R'isa free statistical software used to analyse the OSNdata. The purpose of this research paper is to explore twitter data related to 'Jammu and Kashmir floods' event, in order to determine the potential use of this data in flood damage assessment. Data generated from Twitter determine the features of information associated to the flood event, the most frequently used words related to flood event and to detect post flood information from Twitter messages.

The data is gathered using the Twitter Search API. We used data visualization analytics supported by 'R' which helps to search, organize, and examine the tweets related to the flood event. Further extracted the information from the tweets related to such disastrous events happening in and around over a period of time. Also inferred the users' activities related to the target event.

The paper is further organized as follows: In the next sections, literature on Twitter data usage, text mining procedures in 'R', experimental evaluation for 'Jammu and Kashmir floods' and at the end conclusions and future works are presented.

.

II. RELATED WORK

In information society, OSNs have become the most important element [3, 4]. Data gathering from OSNs is a difficult task, but with the advent of APIs provided by the social media sites it is easy to retrieve data from OSNs. 'Twitter' the most popular OSN is a source of information used to predict the occurrence of disastrous events and also helps the relevant users in obtaining disaster related information [5]. 'Tweedr' tool phases: classification, clustering and extraction, are used to analyze the tweets related to 12 different crises in the United States disaster damage or causalities and specific information about different classes of infrastructure damage, damage types and casualties since 2006 [6]. Use of Twitter in extracting the tweets during the Australian 2010-2011 floods and the role of online communities for the Queensland, New South Wales and Victorian floods were presented by [7]. The geo-referenced Twitter data related to flood event from all over the world for 69 days from a total of 8242 unique tweet users were collected to assess the damage is presented by [8]. The institutional use of Twitter on the number of followers, users followed and tweets published along with the Tweet stats statistical information on Twitter usage, together with a summary of

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

the average number of tweets posts per month is specified in [9].

'R' is an open source statistical tool is one of the most versatile statistical computing environments available over the years. Information about the users from the tweets over a period of time can be extracted using the text mining. 'R' provides tm package for text mining [10].

It provides the necessary infrastructure to organize, transform, and analyse textual data. Also a survey on various text mining facilities in 'R' is presented in [11]. Twitter APIs can be accessed using user credentials via Open Authentication (OAuth).

These APIs are used to obtain data from Twitter. APIs used to access Twitter data are classified as REST APIs and Streaming APIs [12]. The package twitter of 'R' statistical software is used to download and analyse the Twitter data related to an event [13].

The search Twitter function is used to extract the information on followers, those following, posts, and hash tags, etc. for this event. The package word cloud [14]is used to obtain the quick visual analysis based on the keywords related to the event that are found in the tweets.

III. EXPERIMENTAL STUDY

Tweets related to 'Jammu and Kashmir floods' are extracted from the Twitter using Twitter REST API by constructing the necessary queries. This API facilitates the search by taking words as queries.

Tweets related to the target event are examined to extract the most frequent words by obtaining the frequency of word's occurrence. The sample tweets are presented in Table1. From the tweets from Table 1 it is observed that most of the tweets are related to post flood information and are in the form of news and photos that showed damage to buildings or infrastructure.

The tweets related to the target event are converted to a data frame and then converted to a corpus. Using the tm package available in 'R' the obtained corpus is processed by changing the letters to lower case, and removing the hyperlinks, punctuations, numbers and stop words. To obtain the relationship between the terms and documents, Term Document Matrix is constructed.

The row in the matrix corresponds to the 'term' and column represents the 'document'. Also each entry in the matrix represents the occurrences of the term in the document. The Term Document Matrix generated from the tweets collected is presented in Table 2. From Table2 it is observed that, the Term Document Matrix is composed of 1950 terms and 1022 documents. It is highly sparse in nature with 99% of the entries being zero.

Barplot is constructed as shown in Fig.1 to identify the frequent words from the corpus and the distribution of their frequencies. The words are ordered alphabetically.

From Fig.1 it is observed that the distribution of word frequencies confirms to the proper characteristic of the target event. A sample of the most frequent words: "Flood", "Jammu", "Kashmir", "relief", "victims", "rescued", "crisis", "Zeenews". "flood affected".

Table1. Sample Tweets

"Govt missing, flood victims on their own" Locals accuse ministers of helping VIPs first, say no authorities on... http://t.co/gc1nVJx5VS
Over 1,25,000 flood-hit people rescued in J&K so far: Eighty-nine transport aircraft and helicopters of Indian... http://t.co/Q7a112JsKA
#ModiMinistryModi asks people to donate generously for J&K flood victims http://t.co/Lm0aaDm7Zq
Google Launches Crisis Map For Flood-Affected Jammu and Kashmir http://t.co/beiyYnm1L7 #gadgets #lordofthenet
RT @ZeeNews: J&K floods: Over 1,27,000 flood-hit people rescued so far http://t.co/s7KgNufIkV

Table 2. TermDocumentMatrix

A term-document matrix (1950 terms, 1022 documents) Non-/sparse entries: 10330/1982570 Sparsity : 99% Maximal term length: 75

The word cloud is constructed from the Term Document Matrix as shown in Fig.2 to visualize the most frequent words related the event. Further the words can be examined by the size of the word which correspond to its frequency and also describes how often it occurs.

The word cloud will only display the most popular words in the tweets related to the event without expressing how these words are related to each other. From Fig.2 it is observed that the word cloud identifies the words "flood", "flood hit", "Jammu" and "Kashmir" and "victims" which validates that the information is presented is for 'Jammu and Kashmir floods'.

Words like "rescue" "damage", "relief" and "flood affected" focuses the post flood activities. Some words "donate", "modi" and "govt" specifies the Indian government support for relief operations. Words like "Zee news", "news" and "google" role and involvement of media in communicating flood related information to the rest of the world.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

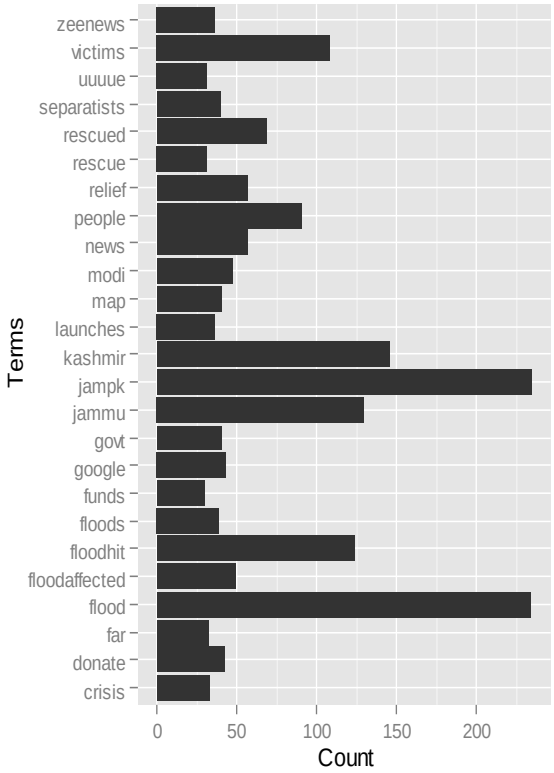

Fig 1. Word frequencies in the corpus

IV. CONCLUSION

In this paper we presented the usage of Twitter and the statistical software R to analyze the Tweets on disastrous event 'Jammu and Kashmir floods'. word cloud is constructed to visualize the most relevant words related the event. Bar plot is constructed which shows the distribution of word frequencies confirms to the proper characteristic of the target event. This analysis will also be helpful in analyzing the various events happening in such related areas over a period of time. Further the future possibilities are to make use of R Hadoop and Mapreduce to create more effective queries to extract and analyze the tweets related to the target event for better results.

Fig 2. Word cloud represents the 'Jammu and Kashmir floods'

REFERENCES

[1] M Goodchild, Crowdsourcing Geographic Information for Disaster Response: A Research Frontier. International Journal of Digital Earth, 2010,vol.3(3), pp. 231-241.
[2] A Bruns, J Burgess, K Crawfordand F Shaw,"#qldfloods and @QPSMedia: Crisis Communication on Twitter in the 2011 South East Queensland Floods", Brisbane: ARC Centre of Excellence for Creative Industries and Innovation, 2012.
[3] C CAggarwal, Social Network Data Analytics: Springer Science Business Media, LLC, 2011.
[4] P D Hoff, A ERaftery and M S Handcock, "Latent Space Approaches to Social Network Analysis", Journal of the American Statistical Association, 2002, vol. 97(460), p.1090–1098.
[5] T Sakaki, M Okazaki and Y Matsuo,"Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors", In Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, p.851–860.
[6] A Zahra, B Christopher, NManojit and C Aron, "Tweedr: Mining Twitter to Inform Disaster Response", In Proceedings of the 11th International ISCRAM Conference,2014.
[7] C France and C Christopher, "Social Media Data Mining: A Social Network Analysis of Tweets During the 2010-2011 Australian Floods", In Proceedings of the PACIS,2011.
[8] M ABrovelli, Z Giorgio, C A Munoz and B Alexander," Exploring Twitter Georeferenced Data Related to Flood Events: An Initial Approach", In Proceedings of the International Conference on Geographic Information Science, Castellón, 2014, p.3-6.
[9] B Kelly B, Institutional Use of Twitter by Russell Group Universities, 2011.
[10] I Feinererand K Hornik,"Package tm: Text Mining Package", 2013.
[11] F Ingo, H Kurtand M David, Text Mining Infrastructure in R, Journal of Statistical Software, 2008, vol.25(5), p.1-54.
[12] S Kumar, F Morstatter H Liu, Twitter Data Analytics: Springer, 2013.
[13] J Gentry, R based Twitter client, 2014.
[14] Ian Fellows, Package 'wordcloud', 2015.