Mining Big Data - An Analysis of SMS Text Data Using R

C. Immaculate Mary
Department of Computer Science
Sri Sarada College for Women (Autonomous)
Salem-16, India
cimmaculatemary@gmail.com

S. Rehna Sulthana
Department of Computer Science
Sri Sarada College for Women (Autonomous)
Salem-16, India
rehnacs@gmail.com

***Abstract*- A communication is said to be resilience when it is effectively used at the time of calamity management. As we are in the global warning era, Natural Disasters are common around the world and help lines are created for victims to communicate with the Disaster Management Systems at the time of emergency. Huge number of peoples who are trapped and affected by catastrophe would try to communicate with an automated or computerized SMS help lines at the state of emergency for rescue, medical or fire emergencies and food supply etc. Humongous and discernable SMS will be generated by the fatalities which leaves the digital map out of BIG DATA. Harnessing this huge and unstructured data will yield several interesting paradigm and valuable information for contemporary decision making and predictive analysis. This paper proposes a frame work for receiving, analysing and visualizing SMS text data and discovering patterns with clustering algorithms. This paper also discusses how the uncovered hidden value becomes serviceable information for current decision making and prophetical study to reduce risks in Disaster Management and Mitigation.**

1. INTRODUCTION

Chennai receives 490 mm rainfall in December 2015 which was a worst rainfall that not eventuates in past 100 years. 500 people were died, 1.8 million people were displaced and 200 million losses recorded at that disaster. In Nagapattinam 12 cyclone shelters were put together and 11 teams from NDRF (National Disaster Response Force) were accelerated to rescue operations. Several toll free numbers and Help lines were announced in affected areas all around South India. An immense number of calls, texts, tweets, posts, comments were generated at the time of disaster (Huiji Gao, 2011). This paper trying to automate the texts formulated at that time of emergencies.

2. CROWD SOURCING

Social media plays an important role at the time of disaster (Lindsay, 2011). Several common people and NGOs are come forward to help the victims. Even though the crowd sourcing media connects the service donors and affected people but fails to construct the proper and centralized bridge between service donors and service needier. And it also fails to propagate geo tag i.e., where the victim actually located and what really needed (Huiji Gao, 2011). The solution is a pre-formatted SMS tags for help lines should be announced for proper services which contains the proper location and number of victims and actual service needed i.e., food, medical emergency or electrical emergency or displacement. Mining such SMS will uncover several interesting patterns which would help for future precautions and current mitigation and management at the time of emergencies.

2.1. Sample Pre-Formatted Helpline Tags

The figure 1 shows a computerized SMS helpline announced at the time of Chennai floods. Similar sample SMS received from various mobile numbers for our mining process. The mobile number and location helps to identify the particular area where the services actually needed and what type of service is needed.

2.2. Unstructured SMS Text

Received SMS text has combinations of characters, numbers, symbols, punctuations and special characters. It is an unstructured text data which needs several steps of pre-processing and effective mining techniques for

unveiling hidden patterns (Ranveer Kaur, 2013). The power full open source R language have several packages which includes pre-processing and mining techniques for automation of the text data. A package called <tm> stands for text mining used in this process for mining the unstructured text data which contains several pre-processing and mining and plotting techniques (Feinerer, 2015).

Chennai Floods: New Computerised SMS Helpline Launched for People in Distress
And here's how people in Chennai can send the text message to the number:
Step 1: Type in the distress SMS number on your phone, i.e., 9220092200
Step 2: Type in your emergency with the key word WATER in the message, for example: WATER 5 family members stranded at house no 5, 3rd street, CIT colony, Kolathur
Step 3: Add your own name and number at the end of the message. Your message should read like: WATER Five family members stranded at house no 5, 3rd street, CIT colony, Kolathur. Sent by Arun: 8734566667

Figure 1. Computerised SMS Helpline Tags

3. TEXT MINING TECHNIQUES

The messages received by MobiliGo software which receives the SMS and directly imports the messages to text files which is easily loaded into R for further text processing. The loaded text files are first pre-processed in several steps and then converted into corpus. Frequently occurred words and association and correlation between frequent words were found out from the corpus in order to identify the relationship between the location and requirements in acquired SMS texts. Quantative analysis of words occurrences and association between words are calculated and plotted for analysis of which area is highly distressed, and which locale need greater number of services.

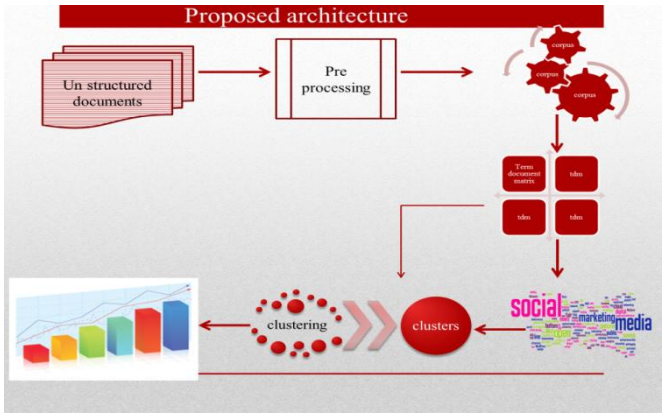

Figure 2. The Architecture of Proposed System

Corpus contained structured text after considerable steps of pre-processing the SMS text. The corpus then converted into term document or document term matrix. This matrix is then used for quantative analysis of words and then clustering and plotting and various visualization techniques such as bar charts, scatter plot and word cloud are used in this framework. The figure 2 shows the architecture of proposed system. In this proposed method, R Studio, one of the power full open source tool has been used for mining the unstructured helpline SMS text. R studio has immense power full packages for data processing. 8500 packages are there and still progressing which may reach up to 10000 packages soon. The packages called text mining<tm>, natural language processing <Nlp>, <snow ballC>, <ggplot2>,<word cloud> are used in this proposed method.

4. MINING SMS TEXT

4.1. Information Retrieval

The messages arrived to helpline number received by the Mobile Go software; this software should be installed in the system and mobile device connected to the system using USB cable, from this software the

messages are directly exported as text files in to the system. Now the text file is ready to load into R studio for processing. The <tm> and <snowballC> Package is used in this processing.

4.2. Pre-Processing

The loaded text file converted as Corpus for text mining. Corpus referred as large and structured set of texts. Corpus has the collection of unstructured text into structured format electronically stored and consisting several documents in it where each document considered being a single record. The corpus contains 88 documents. The text in the documents may contain combination of numbers, uppercase letters, special characters and symbols that should be pre-processed by several steps using the functions available in the text mining package (Graham.Williams, 2016).The contents of corpus converted to lowercase and numbers, punctuations and stop words are removed. R studio contains 174 predefined stop words to be removed and also we can remove our own stop words. After white spaces are stripped and the document was changed as plain text document and then stemmed. Corpus contains many words which have common roots, for example, “trap”, ”trapped”, “trapping”. Stemming is the process of reducing the ends of the words. The above word is stemmed as “trap”.

4.3. Analysis: Conversion of TDM and DTM

The plain text document is then converted to term document matrix or document term matrix. This process is known as conversion of the corpus text into mathematical objects for quantative analysis. The rows and column of Matrix is terms and documents and the cell refers the number of occurrences. Document term matrix represents the relationship between terms and documents, where each row stands for a document and each column for a term, and an entry is the number of occurrences of the of the document term matrix.

4.4. Finding Frequent Items

Frequently occurred items are found out using find FreqTerms method from the TDM. Words occurred frequently for minimum 5 times are displayed in figure 3 and they are ordered alphabetically. Figure 4 shows the frequently occurred words are which bar plotted for visualization.

```
[1]  "adayar"      "ambul"        "boat"      "download"    "emerg"
[6]  "food"        "httpbitlyway" "medic"     "member"      "near"
[11] "need"        "packet"       "peopl"     "perumbakkam" "requir"
[16] "rescu"       "saidapet"     "second"    "send"        "sent"
[21] "street"      "ten"          "trap"      "twenti"      "urgent"
[26] "velacheri"   "via"          "want"      "water"       "waysm"
```

Figure 3. Frequently Occurred Words

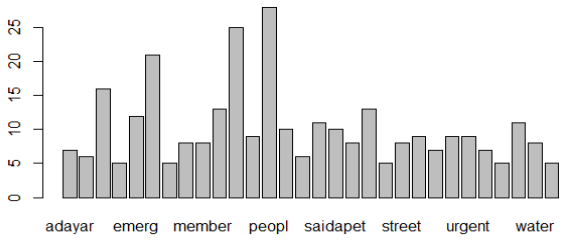

Figure 4. Bar plot of Frequently Occurred Words

4.5. Association between Words

Table 1 and 2 shows association between words with 0.3 correlation limit. Correlation is the measurement of association between two words. If the words are not correlated, the measurement is 0.0 and they are not associated with each other. From this association the word “ambulance” is revealed as a pattern which is highly associated with the areas little mount and Perumbakkam. The term “boat” highly associated with Anna agar and Taramani. The word “food” highly associated with the areas Adayar and Poonthamalli. Water bottles greatly required in Pallikaranai, Poondamalli, Sowkarpet. Rescue boats needed for the areas Anna agar, Taramaniand, Tnagar. And the association between locations and service also represented with 0.3 correlation

limit. Adayar area highly needs the food packets. There is an emergency need of electrical and medical service in Perumbakkam. Clothes needed for valechery area

Table : 1 Association between Service and Location

Ambul	Boat	Food	Water
littl 0.56	rescu 0.80	packet 0.60	bottel 0.69
mount 0.56	trap 0.62	adayar 0.53	pallikaranai 0.59
urgentlti 0.56	replac 0.46	chennai 0.39	poontham 0.59
urgent 0.50	annanagar 0.40	chennai 0.39	second 0.45
perumbakkam 0.47	taranani 0.40	poontham 0.34	sowkarpet 0.34

Table : 2 Association between Location and Service

Adayar	Perumbakkam	Saidapet	Velacheri
chennai 0.54	electrit 0.61	trapp 0.61	cloth 0.64
food 0.53	pleas 0.61	twenti 0.59	flood 0.64
need 0.37	ambul 0.47	littl 0.43	requir 0.42
urgent 0.32	urgent 0.47	mount 0.43	packet 0.32

4.6. Frequency Plotting

The plain text document is converted to document term matrix using the Document Term Matrix method. Terms from each and every document are calculated using column sum function. The table 3 shows frequency of terms in total documents and their frequency is ordered and plotted from the table the frequently occurred locations and needed items are plotted in figure 4, 5, 6.

Table : 3 Ordered Frequently Occurred Words

Terms	Freq.	Terms	Freq.	Terms	Freq
approxim	1	bottel	4	medic	8
merina	1	chennai	4	member	8
life	1	cross	4	second	8
shelter	1	electr	4	street	8
sowkarpet	1	electrit	4	water	8
stay	1	koyammedu	4	packet	9
anna	2	pleas	4	ten	9
five	2	replacc	4	twenti	9
littl	2	tnagar	4	urgent	9
mount	2	trapp	4	perumbakkam	10
nagar	2	download	5	saidapet	10
urgentlti	2	httpbitlyway	5	rescu	11
affect	3	sent	5	want	11
annanagar	3	via	5	emerg	12
cloth	3	waysm	5	near	13
flood	3	ambul	6	send	13
pallikaranai	3	requir	6	boat	16
poontham	3	adayar	7	food	21
salem	3	trap	7	need	25
taranani	3	velacheri	7	peopl	28

From this frequency plotting the study shows that, highest number of SMS is generated from the areas such as, Vela Cheri, Peumbakkam and Saidapet and the greater number of services required is water bottles, food and rescue boats and medical emergency. These details are plotted separately using bar plots. From this study the mitigation management system can be alerted for which area need highest service and which service is highly required. This is not only for current situation and also used for future precautions.

4.7 Word Cloud

Word cloud is generated for instant visualization of frequently occurred words by using word cloud package and its function. The frequently occurred word has bold and large visualization.

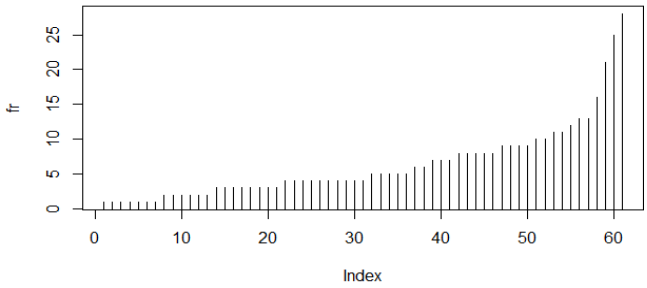

Figure 5. Frequency Plotting of Ordered Frequent Words

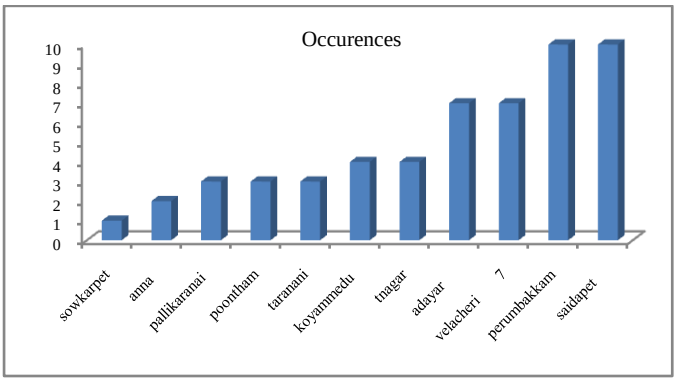

Figure 6. Bar Plot of Locations and Their Occurrences

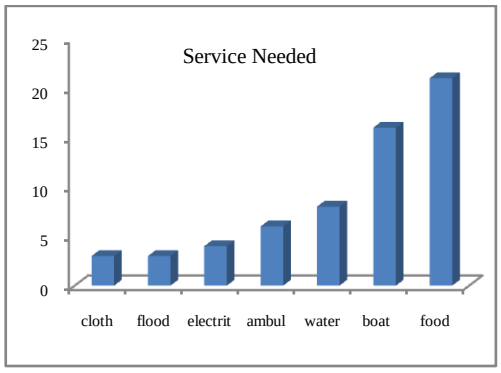

Figure 7. Bar Plot of Needed Service and Their Occurrences

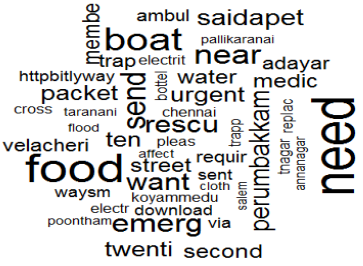

Figure 8. Word Cloud of Frequently Occurred Words

4.8. Clustering

4.8.1 Hierarchical Clustering

The term document matrix has 93% of sparsity, for clustering the sparsity should be removed. The distance between the terms is calculated using distance scaling method. And the terms are clustered according to its occurrences using hierarchical clustering method. Figure 7 shows the agglomerative ward method is used in this clustering and the dentogram is divided into 6 clusters.

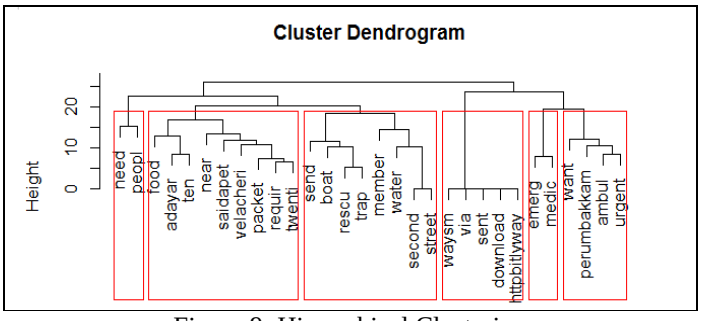

Figure 9. Hierarchical Clustering

4.8.2 k-means Clustering

Clustering is grouping up of similar objects into different groups. K-means is one of the standard clustering algorithms which group the similar objects into given number of clusters. The term document matrix is clustered after removing sparsity of matrix with sparse value 0.95. Then the matrix is converted into proper

matrix. The number of clusters is 6 and centroids are selected randomly with set seed function. The resultant clusters of words are displayed with each 8 words from each cluster in the figure 9.

```
cluster 1: boat need member peopl second street water download
cluster 2: emerg medic perumbakkam send near want adayarambul
cluster 3: peopl need food ten adayar boat rescu trap
cluster 4: perumbakkam urgent want ambul food packet adayar boat
cluster 5: food near packet peoplrequirtwentivelacherisaidapet
cluster 6: ambulemerg medic need saidapetadayar boat download
```

Figure 10. k-means Clustering

5. ANALYSIS AND PROPHETICAL STUDY

Several interesting patterns revealed after analysing the sample SMS received from different locations. The frequently occurred words show that the location from where highest SMS are received and which service is highly needed. The words Perumbakkam and Saidapet occurred 10 times, so it shows that SMS from this locale is highly received. The words food, medical emergency, boats occurred frequently it shows that which service is highly needed. The association of terms shows that, which service is needed for which area. This text analysis may not produce an accurate result but reveals some approximate conclusions that which locality should be alerted with rescue boats and ambulances and for which places the calamity management system should send more food packets and water bottles.

6. CONCLUSION AND FUTURE WORK

The above study reveals some hidden patterns which is very useful for future precaution steps and mitigation management. This process can be enhanced in three levels ie, receiving big uncertainty data, storage and processing. First the unstructured data can be taken from various sources like texts, tweets, comments, posts, what's app texts form social media. Secondly, to store this bulk amount of texts, we can use the power full HADOOP, which is concerned for big data. Rhadoop and Rhipe open source tools are available for integrating R with Hadoop. Thirdly ,the processing can be enhanced using some data mining techniques such as, Association rule mining, Classification of SMS based on location from which locale they are received, clustering the term document matrix using standard K-means techniques and reducing the sparsity of the matrix.

REFERENCES

Feinerer, I. (2015). Introduction to the tm Package Text Mining in R - R Project. Retrieved from cran.r-project.org/pub/R/web/packages/tm/vignettes/tm.pdf

Graham.Williams. (2016). Hands-On Data Science with R. Retrieved from Data Science with R Text Mining - OnePageR - Togaware: onepager.togaware.com/TextMiningO.pdf

Huiji Gao, G. B. (2011). Harnessing the crowdsourcing power of social media diaster relief. (D. Zeng, Ed.) Ieee InTeLLIGenT SySTemS, 14.

Lindsay, B. R. (2011). Social Media and Disasters: Current Uses,Future Options, and Policy Considerations. Congressional Research Service. American National Goverment: Congressional Research Service.

Ranveer Kaur, S. A. (2013). Techniques for Mining Text Documents. International Journal of Computer Applications, 66, 29.