Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 5 – No.6, July 2016 – www.caeaccess.org

Understanding 2016 Drought in India by Social Media Data Mining

Vaishali J. Shimpi
M.E student, Department of C.E.
Dr. D.Y. Patil School of
Engineering & Technology, Pune

Roshani Raut (Ade)
Associate Professor
Dr. D.Y. Patil School of
Engineering & Technology, Pune

ABSTRACT

Drought is natural Biohazard, which result due to deficiency of rainfall for consecutive years. It can last for a few months to a year. It produces short term to long term impact on human society. Understanding of impact of drought can provide great help for disaster management and rehabilitation, It can also provide a way to understanding society and it issue. Social media data mining can aid more effectively and faster study of drought. This paper puts focus on the impact of drought in India 2016.

Keywords

Drought, Social media, Data mining, Classifier, Radiant6, Navies Bays, NodeXL.

1. INTRODUCTION

Drought is Biohazard of nature. It is referred as a "creeping phenomenon" and its impacts vary from region to region.[1] In a general sense, drought result from a deficiency of precipitation over an extended period of time, usually a season or more resulting in a water shortage for some activity, group, or environmental sector. Its impacts result from the interplay between the natural event (less precipitation than expected) and the demand people place on water supply, and human activities can exacerbate the impacts of drought. Because drought cannot be viewed solely as a physical phenomenon, it is usually defined both conceptually and operationally. To understand and true impact of drought is always a challenge. Here this paper will demonstrate how social media data mining can help to understand the impacts.

In India main water source is monsoon rainfall and from 2013 India got below average rainfall, which lead to severe water deficiency in many regions which resulted drought in 2016. They're many ways used to understand drought impact and action for rehabilitation of drought affected area by government agencies, NGO and private organization. Social media data mining can help to understand the impact and rehablitaon work.

Social media is “ A group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content.” [2] Wide availability of the internet people are using social media in their day to day life for interacting and stay connected. People always like to express and share their views and comments for anything they do and see in their life with another. Social media provides that way to be connected with each other though they are far and not able to see each other.

Social media are generating huge data for person’s expression which is more spontaneous and generic to any issue. This data is readily available on the internet, which can be used to understand the people view and expression to a particular issue and this understanding can be used to fix that issue. Some time solution to the issue is also shared by people and this solution can more realistic and close to people. Policymaker, researcher and data analyzer can use this data to form new policies and understand trends and issue in society.

Social media data are very huge which make it very complicated to segregate a particular data set and with no firm rule and format. This available data always very noisy and remove that noise from data is always challenging to data analysis and research. In this paper, discuss the method used by some researcher for their study and comparison between them.

2. RELATED WORK

2.1 Research on: Learning on Engineering Student Issues

In this research, researcher done work towards learning engineering student issue. Researching used twitter as his social media for collecting data. Then he used different manual as well as automated noise removal techniques and establishing different keyword towards issue with the help of multi label Navies Bays classifier to classify data. This is used to understand learning on the engineering student issue.

For Data collection in this reasearch researcher used Twitter as social media data source. As the data collected from twitter are very noisy this data need to be preprocessed by inductive analysis. As data from twitter has largely of informal language, acronyms, sarcasm, and misspellings, meaning is mostly ambiguous and subject to human interpretation. This social media data is if directly feed to automatic algorithm,it may generate fails result. [3] So inductive analysis is a most need activity in which data is revoked for quality. For that researcher manually taken some sample data and analyzed for prominent category formation. Then check the quality of data and arrive at the best possible solution the sample data is inter rated by three scientists. In that one tweet is rated by three scientists for arriving at conclusions which category that should go. But it was observed by them that one tweet can be labelled by two or more categories. So by using F1 measures [4] harmonic mean of two sets of data is extracted. And this way researcher arrived with six categories in which data can be labeled.

Inductive analysis showed that multiple classifiers are most useful for making automated classification data collected by researchers. One of the popular methods of using multi-label classifier is to divide data in multiple single labels. [5] On single label data, binary transform can be used to make the classification effective.

2.2 Research on: Social Media In Disaster Response

This research is focusing on the use of social media data for disaster management with a case study for Queensland flood and associated response. There are many possible ways to targeting reposes based on communicating objects. This research is done on the responses between Emergency Respond agencies and the general public. [6]

Here data collection is done by Queensland Police by creating pages for department on Facebook, which are used to connect to public. General public makes micro-blogging on this page regarding issues. So, Data is collected manually checking the each micro-blog. On which genre analysis is done.

Genre is a particular type of category in association of literature or art. [7] In Genre Analysis we are segregation content or micro blog into different genres. Analysis performed as per following step.

Step 1: Allocating specific genre of micro blog

Step 2: Analysis of genre to reduce it for the Top level genre.

Step 3: Analysis of top level genre and confirm performers of micro blog

Step 4: Mapping genre in chronological order

Step 5: Understanding minimum and maximum impacting genre.

This way data set is analyzed and social media data can be converted into meaningful output.

Above related work toward social media data mining helped in understanding the different social issue.

3. TWITTER DATA

Twitter is one of most prominent social media desk were many people express their view with very short word. Twitter has its own language to communicate like to show subject relevance different word starting with # is used which is called as **Hashtag** example, are #NarendraModi, #College… etc. Same way if someone reply or expresses his view on another it start with RT@ this common practice followed on twitter make it unique and which inherently attract data analyzer for study for it for different topics.

This way many researchers like Gaffney [8] who analyzed tweets with hashtag #iranElection[9] to quantify online activism of the user and presented using histograms, user networks, and frequencies of top keywords. Similar studies have been conducted in other fields, including healthcare [10], marketing [11], athletics [12], just to name a few. Analysis methods used in these studies usually include qualitative content analysis, linguistic analysis, network analysis, and some simplistic methods such as word clouds and histograms. Same way here through this paper another approach to collect data from twitter is proposed.

4. DATA COLLECTION

To collect data from twitter many ready tools are available such as Radian6, NodeXL etc.[13] In this study ready available API of twitter is used with very little modification. This way the data for drought in India is collected. Base on study topic different hash tag searches for here https://ritetag.com/ is used to get some major hash tag some examples as shown below.

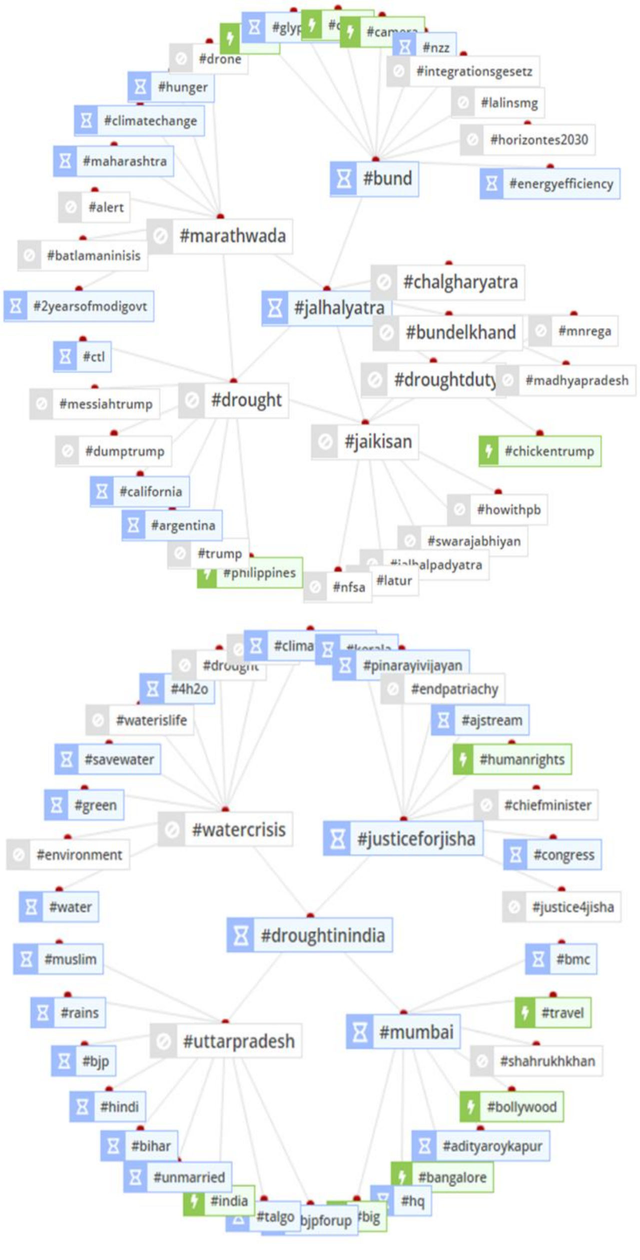

Fig 1: drought related tweets using NodeXL

And twitter related to this hash tag is collected by twitter API . For current study, 100 tweets are collected and studied.

5. DATA PREPROCESSING

As discussed twitter data are very noisy due to its own way of communication many times this data has some symbols an many repetitions of words and sentences. Before going for any analysis this data need to preprocessed to eliminate noise present in the data. For understanding that for drought a algorithm is used to preprocess data. Flow diagram for algorithm is shown below. By doing preprocessing 100 tweet reduced to 72 tweets.

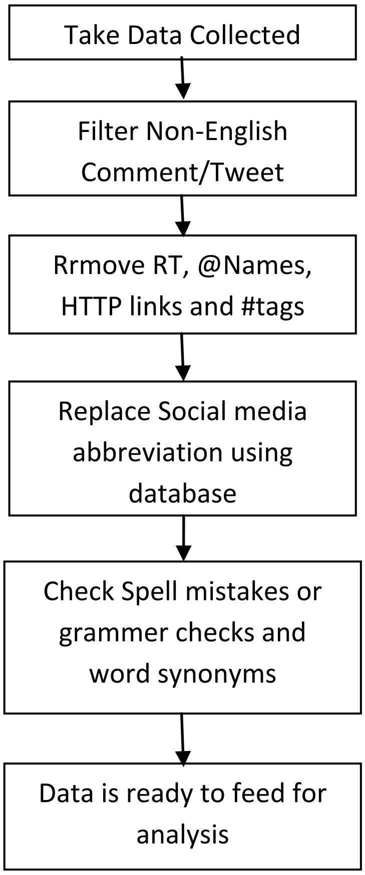
Take Data Collected
Filter Non-English Comment/Tweet
Rrmove RT, @Names, HTTP links and #tags
Replace Social media abbreviation using database
Check Spell mistakes or grammer checks and word synonyms
Data is ready to feed for analysis

Fig 2: Flowchart for data preprocessing

6. DEVELOPMENT OF CATEGORIES

To form a categories data study carried out and study of different literature is done. By this mainly three aspect came in picture issues that are caused by drought, Actions carried out track drought and government reaction in drought. These major categories are again associated with keywords to make an analysis.

6.1 The issue caused due to drought

With the drought, many issues surfaced like due to water shortage no cultivation happens which lead to food price increase as well as lead food deficiency. No work due industry closing and labor, immigration from one place to another. Based on study, some keyword identified which can come under this categories like Water, dry, job, river, rain, problems, worst, heat.

6.2 Major Action

To tackle above issues many actions are taken like water delivery, Food Shelter for human and animals, work projects etc. based on data study we extracted keyword for this category is tanker, relief, relax.

6.3 Government Reaction

Government takes some actions to tackle drought and rehabilitate effected area such as providing water resource, work project.etc based on study keyword associated with this category is govt, minister.

7. METHODOLOGY FOR CLASSIFICATION

Based study and referring other related work for classification of data Naïve bayes classifier is choosed for classification

7.1 Naive Bayes Multi-Label Classifier

The Naïve Bayes algorithm is one of the most important supervised machine learning algorithms for classification. This classifier is a simple probabilistic classifier based on applying Bayes' theorem as follows:

$$\mathrm{P(A|B)} = \frac{\mathrm{P(B|A)P(A)}}{\mathrm{P(B)}}$$

Naïve Bayes classification has an assumption that attribute probabilities $P(X_i|C_j)$ are independent given the class C_j, where X_i is i^{th} attribute of the data instance. This assumption reduces the complexity of the problem to practical and can be solved easily. Despite the simplification of the problem, the Naïve Bayes classifier still gives a high degree of accuracy.[14,15,16]

7.2 Implementation

Proposed data set from twitter for drought D and dj is random tweet d϶D and it check against label sat L where l϶L. By Naïve Bayes for multipliable classifier : H(d)=d→{l, ¬l} Each tweet is compared for each label for its adherence and it is classified.

Let d_j is random tweet , it's feature is ($f_{1,}f_{2,\dots .}f_{m),}$ here f_g is gth feature in tweet, Document label set is $L=\{l_{1,}l_{2,}\dots .l_n\}$. The conditional probability of document $d_{j.}$ with related to each class label $P(l_i|d_j)$ is defined as follows with Naïve bays theyrom.

$$\mathrm{P}(\mathrm{l_i}|\mathrm{dj}) = \frac{\mathrm{P}(\mathrm{d_j}|\mathrm{l_i})\mathrm{P}(\mathrm{li})}{\mathrm{P}(\mathrm{d_j})}$$

As individual probability of label $P(l_i)$ does not change result, it can be ingnored. And probability of tweet adherence to preticulaer label $P(l_i|d_j)$ can be obtained from following formula.

$$\mathrm{P(di|li)} \approx \prod_{g=1}^{m} \mathrm{P(fg|li)}$$

Here, $P(l_i)$ and $P(f_g|l_i)$ can be estimated according to following formulas

$$\hat{\mathrm{P}}\ (\mathrm{L{=}l_i}){=}\frac{\mathrm{N_i}}{\mathrm{N}}$$

$$\hat{\mathrm{P}}\ (\mathrm{fg}|\mathrm{l_i}){=}\frac{1+\mathrm{Ngi}}{\mathrm{m}+\sum_{g=1}^{m} N_{gi}}$$

Here N_i is the amount of tweet having the lebal l_i. N_{gi} is the total frequency of word f_g appearing in tweet data in category l_i. For single-label classifier, the predicted category of document dj is the maximum probability of li categories. For

Multi label classification, for representing posterior probability of tweet dj in each category.

$$\mathrm{P}_{\text{multi}\,\_\,\text{label}} = \frac{1}{n}\sum_{i=1}^{n} P(li|d_3)$$

This way, when $\mathrm{P}_{\text{multi-label}} \leq \mathrm{P}(l_i|d_j)$, dj will labeled as li. Based on this naïve bayes muti label can be as follows

$$\mathrm{H(d)} = \bigcup_{l_i \in L}\{li: P(li\,|\,\mathrm{d}) \geq \mathrm{P}_{\text{muti-label}}$$

Above equation give least probability of particular belong to a particular category or label and this way that particular tweet is label on the maximum probability label which it has.

7.3 Experimental Result

As discussed in below table based on study categories and associated keyword's shown

Table 1: Categories and keyword

Category	Keywords
Issue	Water,Dry, Job,River, rain, Problems, worst,heat
Measure plan	Tanker, relief , relax
Govt. Action	Govt, minister

And by applying a Naïve Bayes classifier for preprocessed data from twitter as per methodology results got is as follows

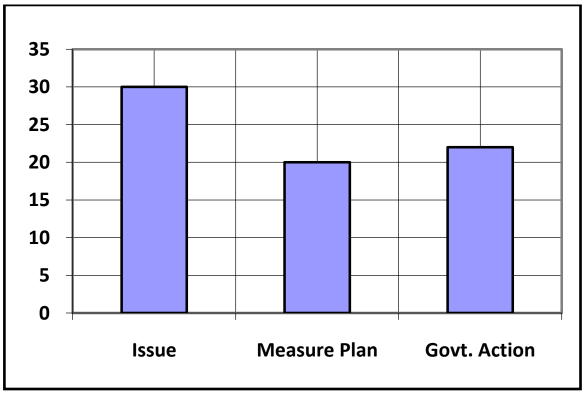

Fig 3:Graph %of tweet vs category

8. FUTURE SCOPE AND CONCLUSION

Finally the workflow proposed requires less human effort for the data analysis and interpretation.There is currently unsupervised automatic natural language processing technique can achieve the depth of understanding. The labels generated can be applied to any similar data sets of problem without extra human effort. Often time spend on analyzing the actual data.

This study puts the lights on Naïve Bayes multi-label classifier in aspect of its efficiency and effectiveness for big data mining dataset from social media.

Advances natural language processing techniques can apply in the future to provide topic recommendation, and further to argument the human analysis results, but a not completely rule out the human effort.

Using this new approach to understanding Drought 2016 in India which provide great help for relevance decision makers to gain further understanding of Drought and provide one way to getting analysis data related different social issue.

9. REFERENCES

[1] Farhood Golmohammadi, Indian Research Journal of Extension Education Special Issue (Volume I), January, 2012 "Drought and it's Environmental and Socioreconomic impacts in the viewpoint of farmers in south Khorasan province-East of Iran".

[2] Kaplan Andreas M., Haenlein Michael (2010) Business Horizons 53, "Users of the world, unite! The challenges and opportunities of social media".

[3] M. Rost, L. Barkhuus, H. Cramer, and B. Brown, Proc. Conf. Computer Supported Cooperative Work, pp. 357-362, 2013,"Representation and Communication: Challenges in Interpreting Large Social Media Datasets".

[4] J.L. Devore, Probability & Statistics for Engineering and the Sciences. Duxbury Press, 2012.

[5] V. Van Asch, "Macro and Micro Averaged Evaluation Measureslys," 2012 http://www.cnts.ua.ac.be/$vincent/pdf/ microaverage.pdf

[6] Christian Ehnis, Deborah Bunker, 23rd Australasian Conference on Information Systems, " Social Media in disaster Response: Queenland Police Service-Public Engagement During 2011Floods.

[7] Maria Jose L, IEEE transactions on professional communication, Vol 48, no3, september2005, "Genre analysis in technical communication".

[8] Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. (2003). Annual Review of Psychology, 54, 547-577. Psychological aspects of natural language use: Our words,ourselves.

[9] D. Gaffney, Proc. Extending the Frontier of Society On-Line (WebSci10), 2010."#iranElection: Quantifying Online Activism".

[10] S. Jamison-Powell, C. Linehan, L. Daley, A. Garbett, and S. Lawson, Proc. ACM Ann. Conf. Human Factors in Computing Systems, pp. 1501-1510, 2012."„I Can‟t Get No Sleep‟: Discuss in #Insomnia on Twitter".

[11] M.J. Culnan, P.J. McHugh, and J.I. Zubillaga, MIS Quarterly Executive, vol. 9, no. 4, pp. 243-259,210 "How Large US Companies Can Use Twitter and Other Social Media to Gain Business Value".

[12] M.E. Hambrick, J.M. Simmons, G.P. Greenhalgh, and T.C. Greenwell, Int‟l J. Sport Comm., vol. 3, no. 4, pp. 454-471, 2010. "Understanding Professional Athletes‟ Use of Twitter: A Content Analysis of Athlete Tweets".

[13] Alemu Molla, Yenewondim Biadgie and Kyung-Ah Sohn, Department of Computer Engineering, Ajou University San5, Woncheon-dong, Yeongtong-gu, Suwon 443-749, South Korea," Network-based Visualization of Opinion Mining and Sentiment Analysis on Twitter".

[14] Xin Chen, Student Member, IEEE, Mihaela Vorvoreanu, and Krishna Madhavan, Xin Chen, Student Member, IEEE, Mihaela Vorvoreanu, and Krishna Madhavan," Mining Social Media Data for Understanding Students' Learning Experiences".

[15] Mehran Amiri, Mahdi Eftekhari, Farshid Keynia, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013,"Using Naïve Bayes Classifier to Accelerate Constructing Fuzzy Intrusion Detection Systems".

[16] Yugang Dai and Haosheng Sun, Journal of Chemical and Pharmaceutical Research, 2014, 6(7): 1636-1643," The naive Bayes text classification algorithm based on rough set in the cloud platform".

[17] Araken M Santos, Anne M P Canuto and Antonino Feitosa Neto, International Journal of Computer Information Systems and Industrial Management Applications ISSN 2150-7988 Volume 3 (2011) pp. 218-227," A Comparative Analysis of Classification Methods to Multi-label Tasks in Different Application Domains".

[18] Zhihua Wei, Hongyun Zhang_, Zhifei Zhang, Wen Li, Duoqian Miao, International Journal of Advanced intelligence Volume 3, Number 2, pp. 173-188, July, 2011," A Naive Bayesian Multi-label Classi_cation Algorithm With Application to Visualize Text Search Results"