A Review: Social Media Data Mining for Understanding and Analyzing Different Issues in Society

Vaishali J. Shimpi
M.E. Student Department of CE
Dr. D. Y. Patil School Engineering & Technology, Pune
Savitribai Phule Pune University

Roshani Ade
Assistant Professor
Dr. D.Y. Patil School ofEngineering & Technology, Pune
Savitribai Phule Pune University

ABSTRACT
Now days social media are becoming day to day part of our life. It can be observed from the number of users increased on social Media site like Facebook, Twitter, YouTube… etc. Social media is a large collection of data for user views and expressions which cannot be ignored by data analyzer, Policymaker, NGO's and organizations based on people view.

Social media data gives very generic and valuable information about problem in society which is studied and problem can be fixed.

Social media data is vast, unstructured, noisy and dynamic in nature, and thus novel challenges arise. This paper reviews some method adapted by researcher to collect, classify and analyzing social media data for understanding some of the targeted issues in society.

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

1. INTRODUCTION
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." [1] 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 view and comment 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 to a particular data set and with no firm rule and format to person's expression mention of social this available data always very noisy. Remove that noise from data is always challenging to data analysis and research. On paper, discuss the method used by some researcher for their study and comparison between them.

2. 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 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 then used to understand learning on the engineering student issue.

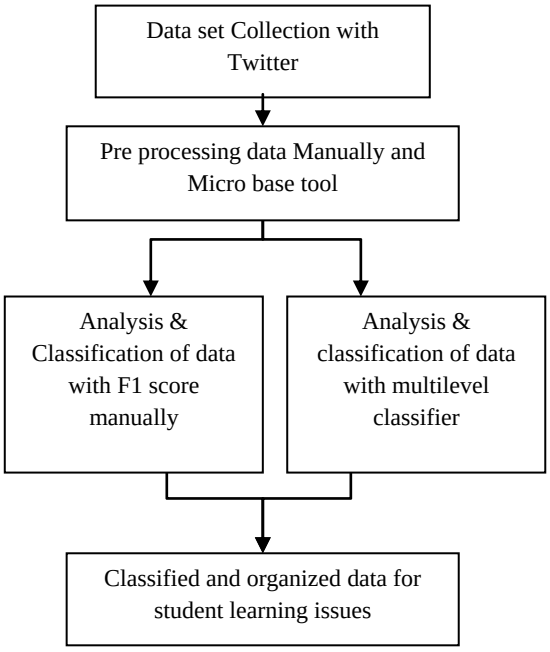

Fig 1: Data Flow Diagram for student learning issue

2.1 Data collection method
Twitter is one of the popular social media to seek help without sharing identity with expert or other people with the same issue. So, researcher chosen twitter as data source.

On twitter, people use # (hash) followed by a keyword followed by an expression as a syntax for communication. (Example -#student) This syntax is called as "**Hash-tag**" in twitter. To collect data on twitter there needs search by specific hash tag. There are many examples of this type of data collection like Gaffney used it for understanding Iran, voters view with hash tag #iranElection.[2] Same way data collection based on hash tag is done in healthcare [3], marketing [4], athletics [5], and it's proven to be very effective.

In this research it started with a random key word like engineer, student, campus, class, homework, professor, and lab, etc., Based on data collected it was found that #engineeringProblems is very reparative and more data is collected by using that. [6]

Data collection Radian6 a commercial data monitoring tool is used.

2.2 Data inductive analysis

Data collected from twitter are very noisy this data need to be preprocessed by inductive analysis. 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. [7] Inductive analysis is a most need activity in which data is revoked for quality. 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. It was observed by them that one tweet can be labelled by two or more categories. Using F1 measures [8] harmonic mean of two sets of data is extracted. 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. [9] On single label data, binary transform can be used to make the classification effective.

3. 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. [10]

3.1 With Facebook data set

3.1.1 Data collection method

Queensland Police already had taken a step by creating pages for department on Facebook, which are used to connect to public. General public makes micro-blogging on this page regarding issues. Data is collected manually checking the each micro-blog.

3.1.2 Data Genre Analysis

Genre is a particular type of category in association of literature or art. [11] In Genre Analysis we are segregation content or micro blog into different genres. Analysis performed as prefollowing 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.

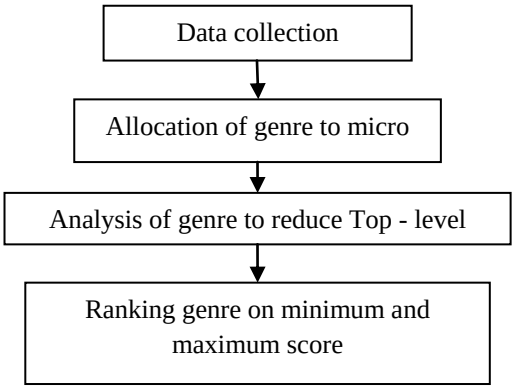

Fig 2: Data Flow Diagram for Disaster response to Facebook data set

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

In the current recherché data collection and genre segregation is manual work its best for a small data set but not suitable for large data set.

3.2 With Twitter data set

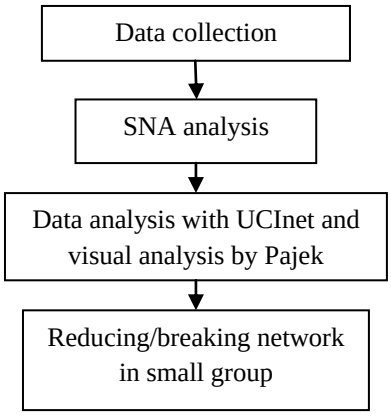

Fig 3: Data Flow Diagram for Disaster response with twitter data set

3.2.1 Data collection method

Collecting data from twitter is possible by using Twitter API or some of the other special tools like Radian6, NodeXL etc., Research purpose researcher done some script based on an algorithm and collected data from twitter.

Challenges for data collection with twitter as follows:

1) Data extraction rate-for is limited to premium member is 350 tweets per hour to 150 per hour for anonymous user.
2) Twitter only allows only 6-day tweet available for public for review, replay and comment.

3.2.2 Data analysis by Social Network Analysis method (SNA)

Social network analysis (SNA) by using a variety of statistical and visual Analysis able to predict the relationship between social member's profiles. SNA used to identify major active group and their activities. This understanding lead to target ting specific group for studying specific subject or issue for group/society. [12]

SNA can be used to check active and inactive of group in social media by which we can predict health of the group.

In this analysis researcher used R and Python scripts for collecting formatted data. UCINet [13] and Pajek [14] are used for data analysis.

3.2.3 Visual network analysis

Visual analysis is based on the graph and link plotted for resources and connections between them. In this analysis major source from the digital communication network is identified to reduce network complexity visually and network is break for small groups to study.

3.2.4 Ego Analysis for user network

Some user is more active and impressive so that they are connected to many users and provide great view keep the group active. This is their ego can be taken and analyzed to understand the issue properly. In disaster management, emergency agencies, volunteer network and some politician can be active an prominent node of a network which help or comment for work or help needed. This study can concentrate around their user network to understand issue depth.

4. RESEARCH ON: AFFECT IN THE WORKPLACE

Organizations always want to gain reputation and earn more profit. It's only possible through effective and more engaged workforce, which work towards organizational goals. Workforce or employee is so much important to organizational that management cannot able ignore employees view and satisfaction. In this research affect of workspace environment and employee view are studies by using social media.

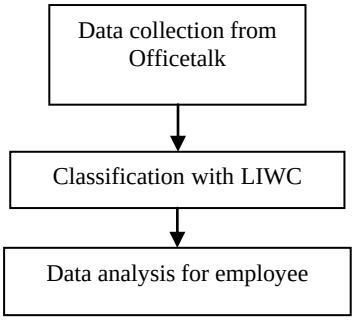

Fig 3: Data Flow Diagram for effect of workspace data analysis

4.1 Data collection method

In office of researcher they use special internal talk application called as Officetalk [15]. Officetalk are same application as twitter. Only differ in number of words allowed. Employee use this tool for micro blogging express their view and frustration on that. Officetalk also store information for user like name, company ID and roll, etc.,

Officetalk are separate organization base software, data easily collected and studied for improving work space for employees.

4.2 Data analysis

In this research to understand employee's real feelings and measurement that is very much important. Researching used prominent psycholinguistic lexicon called as Linguistic Inquiry and Word Count (LIWC) [16]. LIWC help in identifying over 64 behavioral and psychological aspects of employment.

Segregation micro-blog on Positive aspect (PA) and Negative Aspect (NA) with the help of LIWC researcher, formed meaning full data which are used to understand employee concern and appreciation.

5. RESULT FOR STUDENT LEARNING

Student learning, research work researcher is able to prepare a meaning full category vise data which are very useful and generic to student problem. Data is extracted from twitter it provides a new way of data base survey for data collection in a shorter span. With research work they able to come up with a point which making stressful to the student. This is more help to education, growth and academic planning.

6. CONCLUSION

Reviewing this work we able to understand how social media data can be used to understand trends and issues related to public. This understanding is going to help society to track and plan solution to issue very well.

With research work we able to understand that there is a need for further study on eliminating preprocess of data manually. Improvement in data processing base with a label is needed.

This research work provides a new way to implementation of it to new horizons.

7. REFERENCES

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

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

[3] 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".

[4] 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".

[5] 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".

[6] Xin Chen, Student Member, IEEE, Mihaela Vorvoreanu, and Krishna Madhavan. IEEE Transactions on learning technologies, July-September 2014, "Mining Social Media Data Understanding Students' Learning Experiences ".

[7] 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".

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

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

[10] Christian Ehnis, Deborah Bunker, 23rd Australasian Conference on Information Systems, " Social Media in

disaster Response: Queenland Police Service-Public Engagement During 2011Floods.

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

[12] Cross, R., Borgatti, S., & Parker, A. (2003). Making Invisible Work Visible. . In R. Cross, A. Parker & L. Sasson (Eds.), Networks in the Knowledge Economy (pp. 261-282). New York: Oxford Press.

[13] Borgatti, Everett & Freeman 2002, "UCINet for windows software for social network analysis" user guide.

[14] V Batagelj and A Mrvar, university of Ljubljana 1997 ~1999 "Pajek-Program for Large Network Analysis".

[15] Munmun Choudhary and Scott Counts, Microsoft research, " Understanding Affect in workspace via Social media".

[16] 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.