Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 6, June 2015, pg.440 – 446

RESEARCH ARTICLE

Recognizing Student's Problem Using Social Media Data

Pallavi Pagare
Department of Computer Engineering, MET BKC Nashik-03
Savitribai Phule University of Pune, Maharashtra, India
pallavi.pagare22@gmail.com

Abstract— Social media sites allow the creation and communications of user created content. Student's comfortable discussion on social media determined into their educational experience, mind-set, and upset about the learning method. Information from such uninstrumented environments can present important data to account student difficulty. Investigate facts from such a social media can be difficult job. It pays notice on engineering student's Twitter posts to recognize problem and difficulties in their learning practices. This thesis proposes a workflow to set together both qualitative research and significant data mining process. First a sample is occupied from student and then qualitative testing conducted on that pattern which is linked to engineering student's educational days. So only tweets associated to engineering student is composed. It is found that engineering students fall upon problems such as strong learning load, not have of social gathering, and sleep problem. Based on this outcome, a multi-label classification algorithm that is Naive Bayes Multi-label Classifier algorithm is applied to categorize tweets representing student's problems. Here we are applying probabilistically based clustering algorithm. We use the well known Kullback-Leibler divergence to measure similarity between uncertain objects in both the continuous and discrete cases, and integrate it into partitioning clustering methods to cluster uncertain objects. The algorithm organizes a detector of learner trouble. This study presents a approach and conclusion that express how casual social media statistics can present insight into student's happening.

Keywords— Social networking, web-text analysis, Education, Social network analysis, Computer and Education, Data mining

I. Introduction

Data mining research has effectively produced numerous technique, tools, and algorithms for managing huge amounts of data to answer real-world troubles. As social media is widely used for various purposes, vast amounts of user created data be present and can be made available for data mining. Main objectives of the data mining procedure are to collectively handle large-scale data, extract actionable patterns, and gain insightful knowledge. Social media sites such as Twitter, Face book, and YouTube present grand place to students to share happiness and struggle, sentiment and tension, and gain social support. On various social media sites, students talk about their everyday encounters in a comfortable and informal manner. This Student's digital information gives huge amount of implicit information and a whole new viewpoint for educational researchers to know student's experiences outside the prohibited classroom environment. This understanding can enhance education quality, and thus improve student employment, preservation, and achievement [1][2]. The vast amount of information on social sites provides prospective to recognize student's problem, but it raises some methodological complexities in use of social media data for educational reasons. The complexities such as absolute data volumes, the miscellany of Internet slangs, the change of locations, and moment of students posting on the web. Pure physical analysis cannot contract with the ever growing scale of data, while pure automatic algorithms cannot capture in-depth significance inside the data [3]

The research goal of this learning are: (1) To show a work of social media information sense-making for educational reasons, combining both qualitative investigation and large-scale data mining techniques. (2) To discover engineering student's casual discussions on Twitter, in order to knows the problem coming into their life.

This Study prefers to focus on engineering student's comments posted on Twitter about their problems in collage life because: (1) Engineering schools and branch have long been stressed with student employment and preservation topics. Engineering graduates comprise a significant part of the nation's potential labor force and have a direct impact on the nation's financial expansion [4]. (2) Based on understanding of students difficulty decision makers can make more knowledgeable conclusions on proper interference that can help students to conquer obstacles in education and help the student to solve the problem. (3) Twitter is a well-liked social media site. Its content is frequently public and very brief that is no more than 140 characters per tweet. Twitter offer free APIs that is used to stream data and allows developers to build upon and extend their applications in new and creative ways. Access Data from Twitter give developers low latency access to Twitter's global stream of Tweet data. To construct a data mining design or are involved in analytics research, the Streaming API is most suitable for such things. Twitter data is in a suitable format for investigation. Twitter's terms of apply for the data are relatively tolerant. It is generally accepted that tweets are public and available to anybody, hence they permit entrée to any account with no need to request for sanction.

A. Social Network Analysis

Social Networks Analysis (SNA), aims at studying relationships between individuals, instead of individual attributes or properties. A social network is considered to be a group of people, an organization or social individuals who are connected by social relationships like friendship, cooperative relations, or informative exchange. Different DM techniques have been used to mine social networks in educational environments, but collaborative sorting is the most common. Collaborative filtering or social filtering is a method of making automatic predictions about the interests of a user by collecting taste preferences from many users [4].

B. Qualitative Analysis

Qualitative analysis is a technique of examination employed in many diverse academic regulation, by tradition in the social sciences, but also in market research and further contexts. Qualitative researchers plan to gather an in-depth understanding of human actions and the reasons that manage such behavior. The qualitative method examines the why and how of decision making, not just what, where, when. Hence, minor but focused samples are often used than huge samples. Qualitative procedures create information only on the particular cases studied, and any more general terminations are only suggestions. Quantitative methods can then be used to look for experimental support for such research theories

II. LITERATURE REVIEW

M. Clark, S. Sheppard, proposed Academic pathways study: Processes and realities. APS consists a series of longitudinal and multi-institutional studies on undergraduate engineering student's learning experiences and their evolution to work. They used various research methods including surveys, structured interviews; semi structured interviews, engineering design task, and small focus groups. The CAEE website presents research briefs from the APS study including topics such as developing identity as an engineer, conceptions of engineering, workload and life balance, and persistence in engineering as a college major and as a career [5].

G. Siemens proposed Learning analytics and educational data mining: Towards communication and collaboration. Learning analytics and educational data mining (EDM) are data-driven approaches emerging in education. These approaches analyse information produced in educational settings to know students and their learning atmospheres in order to notify institutional management [6].Two research communities - Educational Data Mining (EDM) and Learning Analytics and Knowledge (LAK) have built-up independently to tackle this need. This study fight for improved and official communication and cooperation among these communities in order to distribute research, methods, and tools for data mining and analysis in the service of budding both LAK and EDM fields.

R. Baker proposed The state of educational data mining in 2009: A review and future visions."Educational Data Mining is an rising regulation, concerned with increasing methods for discovering the distinctive types of statistics that come from learning settings, and applying those methods to well again recognizing students, and the situations which they gain knowledge of in." Educational data mining methods have enable researchers to model a broader range of potentially relevant student attributes in real-time, including higher-level constructs than were previously possible[7].

M. Vorvoreanu proposed Managing identity across social networks, Human identity is complex and multifaceted. Identity is acquire through social interaction and enact different roles, The facet of identity one enacts at a given point in time depends upon context and the particular social group i.e. family, coworkers, friends present in that context[8].

III. METHODOLOGY

We built-up a workflow to put collectively both qualitative investigation and large-scale information mining methods it paying attention on engineering student's Twitter posts to know problem and troubles in their educational practices.. Primarily a sample is taken from learner and then it carry out qualitative investigation on that sample which is related to engineering students learning life. It found engineering students encounter problems such as heavy learning burden, not have of social gathering, and sleep insufficiency. Stand on these outcomes, authors apply a multi-label classification algorithm to categorize tweets presenting student's problems. Then used the algorithm to prepare a detector of student problems. This study presents a approach and conclusion that express how casual social media information can present approaching into student's occurrence[2]. In this study it implemented a multi-label classification model where we permitted one tweet to go down into many categories at the same time. Our categorization is compared with other generic classifications. Our work expands the range of data-driven approaches in teaching such as learning analytics and educational data mining.

The important point in proposed study are, First, it propose a workflow to bridge and integrate a qualitative research methodology and large scale data mining techniques. It base our data-mining algorithm on qualitative insight resulting from human understanding, so that it can gain deeper understanding of the data. Then apply the algorithm to another large-scale and unexplored dataset, so that the physical method is improved. Second, the paper provides deep insights into engineering student's educational experiences as reacted in informal, uncontrolled environments. Many issues and problems such as study-life balance, lack of sleep, lack of social engagement, and lack of diversity clearly emerge. These could bring awareness to educational researcher, policy-maker.[1]

A. System Flow

The main goal of the system is to explore engineering students' informal conversations on Twitter, and classify the tweets based on the categories developed in content analysis stage in order to understand issues and problems students encounter in their learning experiences. Tweets is classified by using naive bayes multilable classifier. Each tweet is classified based on their probability values.

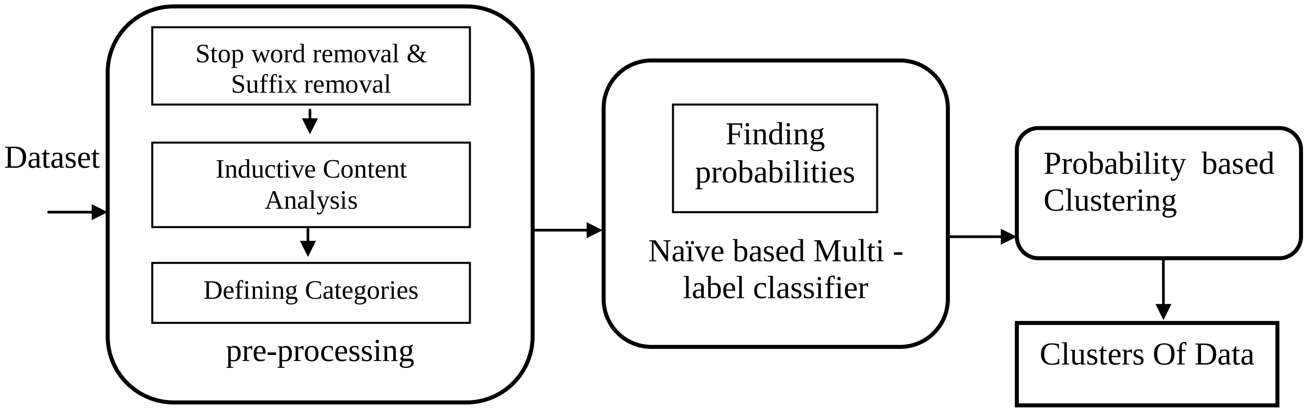
Dataset
Stop word removal & Suffix removal
Inductive Content Analysis
Defining Categories
pre-processing
Finding probabilities
Naïve based Multi - label classifier
Probability based Clustering
Clusters Of Data

Fig.1: System Architecture

Fig.1 shows the system Flow diagram. Twitter Dataset is given as an input to the system. This system consist of an investigative process to find the related information and relevant Twitter hashtag ,a Twitter hashtag is a word start with a #sign, used to emphasize or tag a topic. Then collection of tweets using the hashtag engineering Problems. After that the Preprocessing is perform on the dataset. Inductive content analysis on samples of the engineering Problems dataset is perform , it has been established that most important troubles engineering students come across their learning occurrences fall into numerous important categories. Based on these categories, it implemented a multi-label Nave Bayes classification algorithm. By finding out the probabilities of each category and tweets the classified output is obtain. Performance of the classifier is evaluated by evaluating it with further state-of-the-art multi-label classifiers. Trained dataset is produced using Naïve Bayes multi-lable classifier System used the classification algorithm to train a detector that could assist detection of engineering students' problems. We use the well known Kullback-Leibler divergence to measure similarity between uncertain objects in both the continuous and discrete cases, and integrate it into partitioning clustering methods to cluster uncertain objects. The results could help educationalist recognize weak students and make judgment on proper interference to retain them.

IV. MATHEMATICAL MODELLING

This study built a multilabel classifier to categorize tweets stands on the categories developed in content analysis phase. There are numerous well-liked classifiers generally used in data mining and machine learning field. It establishes that Nave Bayes classifier to be very efficient for this dataset compared with further multilabel classifiers.

A. Pre-processing

Twitter client use various unusual symbols to express certain significance. For Ex, # is used for hashtag, @ used for a user account,RT show a retweet. Twitter users occasionally duplicate letters in words thus to highlight the words, for ex, "soooo cuuuteeee" and "Looooking awesome". In addition, common stopwords such as "a, an", nonletter symbols,punctuation carry noise in text. So we preprocessed the texts prior to training the classifier

- First remove every engineering Problems hashtags. And for new occurring hashtags, just removed the sign, and reserved the hashtag texts.
- For identifying negative emotion and issues negative words are used. thus it replace words finishing with "n't" and further frequent negative words (e.g. no, not, nothing) as negtoken".
- Detached every one word that includes non letter symbols and punctuation. This incorporated the deletion of @ and http links. Also delete all the RTs.
- For replicated letters within words, policy when it discovers two matching letters replicating, it reserved both of them. If it identified more than two same Letters replicating, substitute them with one letter. Therefore, "soooo cuuuteeee" is corrected to "”So cute”. Initially accurate words such as "Sweet" and "buddy" were kept as they were.
- Eliminate the frequent stopwords. It kept words like "much, lot, many, all, forever, still, just", because the tweets regularly use these words to communicate point.

B. Naïve Bayes Multi-label Classifier

The Naive Bayes classifier is a straightforward probabilistic classifier which is based on Bayes theorem with strong and naïve self-government assumptions. Naive Bayes classifier is extremely efficient since it is less computationally and it requires a small amount of preparation information. One well-liked way to execute multi-label classifier is to convert the multi-label organization problem into multiple single-label categorization problems [9].

Assume there are sum of W words in the preparing document compilation in this case, every tweet is a document.D = d1;d2; : : : ; dw, and a total amount of M categories K = k1; k2; : : : ; kM. If a word dw appears in a category k for $n_{d_w k}$ times, and appear in categories other than k for $n_{d_w k}{}'$ times, then, the probability of this word in a definite category c is

$$p(d_w|k) = \frac{n_{d_w k}}{\sum_{w=1}^{W} nd_{w k}} \qquad (1)$$

Similarly, the probability of this word in categories other than c is:

$$p(d_w|k') = \frac{nd_{w k}}{\sum_{w=1}^{W} nd_{w k}} \qquad (2)$$

Suppose there are an entire number of X documents in the preparing set, and K of them are in category k. Then the probability of category k is:

$$p(k) = \frac{K}{X\prime} \qquad (3)$$

And the probability of other categories k’ is :

$$p(k') = \frac{X-K}{X} \qquad (4)$$

For a document di in the trying set, there are Y words Wdi = wi1;wi2; : : : ; wiY, and Wdi is a subset of D. The reason is to classify this document into class c or not c. We imagine independence between all word in this document, and any word wik conditioned on k or k' follows multinomial distribution. Therefore, according to Bayes Theorem, the probability that di fit in to category k is

$$p(k|d_i) = \frac{p(d_i|k) \cdot p(k)}{p(d_i)} \alpha \prod_{y=1}^{Y} p\big(w_{iy}\big|k\big)\, p(k), \qquad (5)$$

and the probability that di fit into group other than c is

$$\mathrm{p}(\mathrm{k}'|\mathrm{d_i}) = \frac{\mathrm{p}(\mathrm{d_i}|\mathrm{k}\prime)\cdot \mathrm{p}(\mathrm{k}')}{\mathrm{p}(\mathrm{d_i})} \alpha \prod_{\mathrm{y}=1}^{\mathrm{Y}} \mathrm{p}(\mathrm{w_{iy}}|\mathrm{k})\, \mathrm{p}(\mathrm{k}') \qquad (6)$$

Because $\mathrm{p}(\mathrm{k}|\mathrm{d_i}) + \mathrm{p}(\mathrm{k}' + \mathrm{d_i})$,it normalize the latter two items which are comparative to $\mathrm{p}(\mathrm{k}|\mathrm{d_i})$ and $\mathrm{p}(\mathrm{k}'|\mathrm{d_i})$to get the actual values of $\mathrm{p}(\mathrm{k}|\mathrm{d_i})$. If $\mathrm{p}(\mathrm{k}|\mathrm{d_i})$ is larger than the probability threshold T, then di fit into category k, otherwise, di does fit into category k. Then do again this process for every category.

C. Uncertain Objects and KL Divergence

This section first models uncertain objects as random variables in probability distributions. We consider both the discrete and continuous probability distributions and show the evaluation of the corresponding probability mass and density functions in the discrete and continuous cases, respectively. Then, we recall the definition of KL divergence, and formalize the distribution similarity between two uncertain objects using KL divergence[10].

1) *KL Divergence*

In general, KL divergence between two probability distributions is defined as follows, Definition 1 (Kullback-Leibler divergence): In the discrete case, let f and g be two probability mass functions in a discrete domain D with a finite or countably infinite number of values. The Kullback- Leibler diverge (KL divergence for short) between f and g is:

$$D(f||g) = \sum_{x\in D} f(x) log \frac{f(x)}{g(x)} \qquad (7)$$

In the continuous case, let f and g be two probability density functions in a continuous domain D with a continuous range of values. The Kullback-Leibler divergence between f and g is:

$$D(f||g) = \int f(x) log \frac{f(x)}{g(x)}\, dx \qquad (8)$$

In both discrete and continuous cases, KL divergence is defined only in the case where for any x in domain D if $\mathrm{f(x)} > 0$ then $\mathrm{g(x)} > 0$ By convention, $0 \log \frac{\mathrm{o}}{\mathrm{p}}$ for any $\mathrm{p} \neq 0$ and the base of log is 2.Note that, KL divergence is not symmetric in general, that is $\mathrm{D(f||g)} \neq \mathrm{D(g||f)}$,

2) *Using KL Divergence as Similarity*

It is natural to quantify the similarity between two uncertain objects by KL divergence. Given two uncertain objects P and Q and their corresponding probability distributions, $\mathrm{D(P||Q)}$ evaluates the relative nncertainty of Q given the distribution of P. In fact, from Equations (7) and (8), we have

$$\mathrm{D(P||Q)} = \mathrm{E}\left[\log \frac{\mathrm{P}}{\mathrm{Q}}\right], \qquad (9)$$

which is the expected log-likelihood ratio of the two distributions and tells how similar they are.

In the discrete case, it is straightforward to evaluateEquation (4) to calculate the KL divergence between two uncertain objects P and Q from their probability mass functions calculated as Equation (2).

In the continuous case, given the samples of P and Q, by the law of large numbers and Equation (6), we have

$$\lim_{\mathrm{s}\to\infty} \frac{1}{\mathrm{s}} \sum_{\mathrm{i}=1}^{\mathrm{s}} \log \frac{\mathrm{P(p_i)}}{\mathrm{Q(p_i)}} = \mathrm{D(P||Q)}, \qquad (10)$$

where we assume the sample of P = {p1, . . . , ps}.Hence, we estimate the KL divergence D(P ||Q) as

$$\mathrm{D(P||Q)} = \frac{1}{\mathrm{s}} \sum_{\mathrm{i}=1}^{\mathrm{s}} \log \frac{\mathrm{P(p_i)}}{\mathrm{Q(p_i)}} \mathrm{c} \qquad (11)$$

To ensure that the KL divergenceis defined between every pair of uncertain objects, we smooth the probability mass/density function of every uncertain object P so that it has a positive probability to take any possible value in the domain.

The smoothing is based on the following two assumptions about the uncertain objects to be clustered,

1. We assume that the probability distribution of every uncertain object to be clustered is defined in the same domain
2. We assume that the domain D is bounded.

We smooth a probability distribution P as follows

$$\mathrm{P(x)} = \frac{\mathrm{P(x)}+\delta}{1+\delta|\mathrm{D}|} \qquad (12)$$

where 0 < δ < 1, |D| is the number of possible values in D if D is discrete and the area of D (i.e., |D| =∫D dx) if D is continuous.

We define the similarity between two uncertain objects as the KL divergence between their probability distributions. The KL divergence is calculated by Equations (7) and (11) in the discrete and continuous cases, respectively.

3) *Clustering Algorithm*

A partitioning clustering method organizes a set of n uncertain objects O into k clusters C1,. . ,Ck, such that Ci ⊆ O (1 ⩽ i ⩽ k), Ci _= ∅,_ki=1 Ci = O, and Ci ∩ Cj = ∅ for any i ≠ j . We use Ci to denote the representative of cluster Ci. Using KL divergence as similarity, a partitioning clustering method tries to partition objects into k clusters and chooses the best k representatives, one for each cluster, to minimize the total KL divergence as below,

$$\mathrm{TKL} = \sum_{i=1}^{k} \sum_{p \in c_i} D(P||C_i) \qquad (13)$$

For an object P in cluster Ci (1 ⩽ i ⩽ k), the KL divergence D(P _Ci) between P and the representative Ci measures the extra information required to construct P given Ci. Therefore,P ∈CiD(P _Ci) captures the total extra information required to construct the whole cluster Ci using its representative Ci. Summing overall k clusters, the total KL divergence thus measures the quality of the partitioning clustering. The smaller the value of TKL, the better the clustering.

V. IMPLEMENTATION STRATEGY

A. *Pre-processing*

For experimentation the tweets from https://twitter.com/engineerproblem site are used for analysis of student problem. From the inductive content investigation stage, we had a total of 45 engineering Problems tweets interpreted with 5 categories. The tweets are applied for preprocessing. The preprocessing stages applied on this data are remove stop words, Removal of hashtag and @ sign , Remove Repeated Characters, Remove HTTP links and Remove non-letter symbol. For Example if the tweeet is "@Ram It's not that it's hard, it's that there's so muchhhhhhhhh info for each class. Wait. No it's both. It's hard and extensive" so after the preprocessing the result will be:

1. Removal of stopword : After this the comment will be "@Ram hard, muchhhhhhhhh info class. Wait. No both. hard extensive." All the stopwords like it, is, that is removed.
2. Removal of hashtag and @ sign After performing this steps the comment will be "Ram hard, muchhhhhhhhh info class. Wait. No both. hard extensive." the @ symbol befor the ram is removed.
3. Remove HTTP links after this step the comment will remain as it is because comment does not contain any HTTP links, so the comment will be "Ram hard, muchhhhhhhhh info class. Wait. No both. hard extensive."
4. Remove non-letter symbol After performing this step the comment will be "Ram hard, muchhhhhhhhh info class. Wait. No both. hard extensive" that is the dot symbol is removed.
5. Removal of double words after this step the comment will " Ram hard much info class Wait No both hard extensive ". "muchhhhhhhhh" is corrected to "much". Initially accurate words such as "class" were kept as they were.

B. *Calculate Probabilities*

We have to First calculate the probability of particular word belongs to specific category C and belongs to category other than C. It is calculated as:-

$$p(d_w|k) = \frac{n_{d_{w^k}}}{\sum_{w=1}^{W} nd_{w^k}}$$

Similarly, the probability of this word in categories other than c is:

$$p(d_w|k') = \frac{nd_{w^k}}{\sum_{w=1}^{W} nd_{w^k}}$$

That is The probability of word in specific category is equal to words appears into that category divided by total number of word. For Ex: Here our comment is "hard much info class Wait No both hard extensive". First we calculate probability of word belongs to category such as "Heavy Study Load" the result of this is 1.0 And probability of word belongs to category other than "Heavy Study Load" the result of this is 1.0.The whole table for Heavy Study load is shown in Table I :

TABLE I
PROBABILITY CALCULATION

C	C Probability	C'	C' Probability
0.0	0.0	1	1.0
1.1	1.1	8	1.0
0.0	0.0	8	1.0

0.0	0.0	10	1.0
0.0	0.0	12	1.0
0.0	0.0	15	1.0
1.0	1.0	8	1.0
1.0	1.0	5	1.0
0.0	0.0	6	1.0

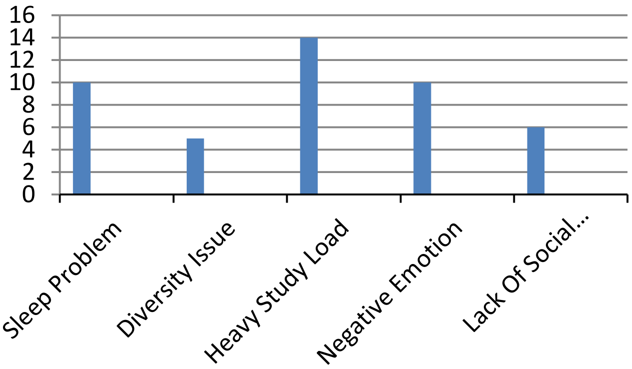

Fig. 2 Graphical representation of tweets classification

VI. Conclusions and Future Scope

Mining social media data is helpful to researchers in Analysis of student's learning Experiences. It gives a way to examining social media statistics. Our study can inform educational administrators, practitioners, decision makers to gain further understanding of engineering students college experiences. It notifies educational manager, and other applicable Stackholder to expand further accepting of engineering students' institution understanding.. Possible future work could analyze students generated content further than texts (e.g. images, videos), on social media sites other than Twitter (e.g. Facebook,Tumbler,YouTube). Future work can also expand to students in other majors and other institutions.

References

[1] Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan." Mning Social Media Data for Understanding Students' Learning Experiences" *IEEE transactions on learning Technologies, ID*, DOI 10.1109/TLT.2013.2296520

[2] Pallavi K. Pagare, "Analyzing Social Media Data for Understanding Student's Problem" *International Journal of Computer Applications* (0975 – 8887), 2014.

[3] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education*," Educause Review*, vol. 46, no.5, pp. 30–32, 2011.

[4] M. Rost , L. Barkhuus, H. Cramer, and B. Brown, "Representation and communication: challenges in interpreting large social media datasets*," in Proceedings of the 2013 conference on Computer Supported cooperative work, 2013, pp. 357–362.*

[5] M. Clark, S. Sheppard, C. Atman, L. Fleming, R. Miller, R. Stevens,R. Streveler, and K. Smith, " Academic pathways study: Processes and realities*," in Proceedings of the American Society for Engineering Education Annual Conference and Exposition*, 2008.

[6] G. Siemens and R. S. d Baker, " Learning analytics and educational data mining:Towards communication and collaboration ",*in Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 2012, pp. 252254.

[7] R. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.

[8] M. Vorvoreanu and Q. Clark, "Managing identity across social networks",*in Poster session at the 2010 ACM Conference on Computer*

[9] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining Multi-label data" *Data mining and knowledge discovery handbook* , pp. 667-685, 2010

[10] Bin Jiang, Jian Pei, Yufei Tao, and Xuemin Lin "Clustering Uncertain Data Based on Probability Distribution Similarity*" IEEE transactions on knowledge and data engineering*, 2011.