

Knowledge Discovery from Social Media Data for Education Enhancement

Ms. Neelam Borde¹, Ms. Neha Dubey¹, Ms. Aakankshi¹

Computer Science and Engineering Department, ICOER, Pune University, Pune, India¹

Abstract: Social media has gained a lot of popularity amongst the students. Students share their feelings and their day to day experiences on it in a very informal and casual manner. Experiences and problems of students revealed through social media need human interaction or human analysis. Knowledge from such uninstrumented environments can present valuable data to report student problem. But mining knowledge from such data can be a very challenging task. The huge amount of data requires automated data analysis techniques. In this paper, a work-flow is developed which combines both qualitative investigation and large-scale data mining scheme. The data posted by students are collected and analyzed. It is found that certain issues like heavy study load, hectic schedule and lack of sleep are encountered by the students. Hence these issues are classified using Naive Bayes Multi-label Classifier algorithm. Here we are also using the clustering algorithm i.e. K-means. This both techniques classification and clustering can help in understanding the student's problem in a very efficient way.

Keywords: Data mining, social media, text mining, social network analysis.

I. INTRODUCTION

There are various types of techniques, tools and algorithms used in data mining for managing huge amount of data. These are used for tackling with real world problems. Social media has become a highly interactive platform to share, co-create, as it depends on mobile and web based technologies. Now-a-days it has become a grand place for students to share their life experiences and gain social support. The students share their opinions in a very casual and informal manner. It has become the easiest and the fastest way of connecting. This Student's digital information gives huge amount of implicit information for educational researchers to understand student's experiences outside the prohibited environment. This understanding can improve the education quality, and thus enhance student employment and achievements. It is evident that social media is important for educational institutes, having an online presence which helps the students to strengthen their connectivity without any limitations. With the rise of social media, the web has become very vibrant and lively. Hence more and more people are actively participating in these sites. Social media has become an ever increasing field in today's world. One to one analysis of the content is not possible for the increasing scale of data, while pure automatic algorithms cannot find the depth of the data. So the research goal of this learning is: (1) To show the sense-making social media information using data mining techniques. (2) To discover the problems faced by the students in their educational life.

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.

B. Qualitative Analysis

Qualitative analysis is a technique of examination employed in much 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. RELATED WORK

G. Siemens et al.* proposed Learning analytics and educational data mining: Towards communication and collaboration. Learning analytics and educational data mining (EDM) comes under data-driven approaches which are now used as a technique in education. These approaches are used for analysing the data produced in educational settings to study the students and their learning atmospheres in order to view institutional management.

Two research communities - Educational Data Mining (EDM) and Learning Analytics and Knowledge (LAK) have built-up independently to solve this need [3].

R. Baker et al.* proposed the state of educational data mining in 2009: "A review and future visions." Mining the Educational data is increasing day by day. They are evolving with distinct a method that comes under learning setting. These methods are again used to solve the problems which student's are facing in their learning life. And it is done in a broader range and potentially [2].

Po-Wei Liang et al.* proposed "Opinion mining on social Media data": This paper focus on micro-blogging data like Twitter, on which user posts their real time opinions about everything, like sharing their feelings. There is much difference between giving reviews about any product and micro-blog messages. The micro-blog messages are short. And the users do not care about any grammar mistakes or any spelling mistakes [4].

Grljevic Olivera et al.* proposed Students' Behavior on Social Media Sites – A Data Mining Approach: This paper focus on analyzing the student's behavior by preprocessing the data. In this CRISP-DM methodology is used. Data was collected through questioner shared among prospect students of Faculty of Economics Subotica.

Clustering techniques was illustrated in this paper to understand the patterns and the profiles of student's behavior.

III. PROPOSE WORK

This paper targets the student's generated data on the site. This raw and unformatted data is taken, classified into various categories and finally clusters are formed. The classification is done based on the words used by the students in their comments and then classified into different classes and finally clusters naming positive response and negative response are made. Navie Bayes and K means algorithm are used for classification and clustering respectively. When we take data from social media then various un-useful data is occurred.

A. Text preprocessing

1. We removed all the symbols like punctuation marks, like @ symbol, etc.
2. Negative emotion is found out by the negative words (no, not, n' t, etc.) in the sentence.
3. The repeated letters in the sentence is removed. It means that if there are two repeating letters than it is made to one .
For e.g. sooo is corrected to so.

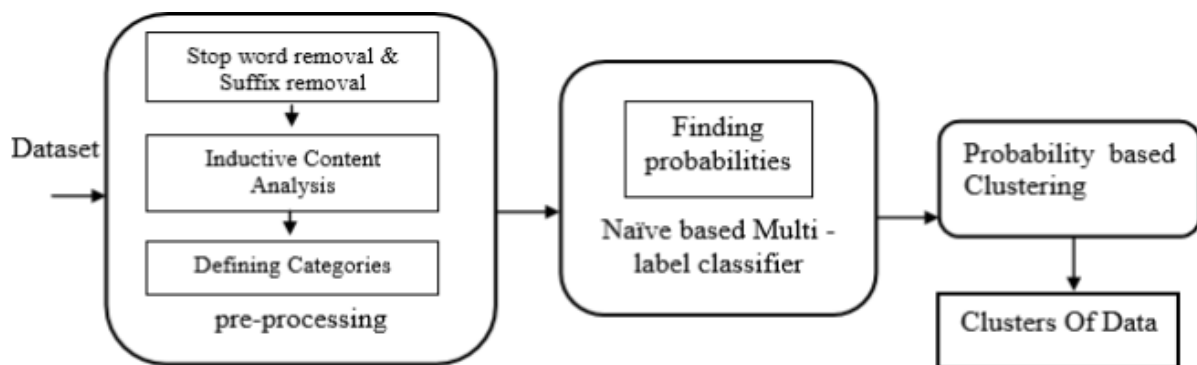


Fig.1: System Architecture

B. Naïve Bayes Multi-label Classifier

The Naïve Bayes classifier is used for the probability classification which is based on the bayes theorem. This classification is basically used for the text mining.

- 1) Suppose there are sum of word say as W, every data is documented as D=d1,d2,d3.....dw.
- 2) Total no. of categories M are K=k1,k2,k3.....km.
- 3) If dw appears in M category k times then probability

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

4) If there are X no. of document in data set and K is the category

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

The probability of other category is,

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

5) The probability of Bayes theorem in category k is:

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

C. K-Means Clustering

k-mean is the unsupervised learning algorithm which is used to solve the clustering problem. In this system we are creating two clusters i.e. positive clusters and negative clusters.

The clustering algorithm is as:

- 1) Define the center K for each cluster.
- 2) It associate with nearest center
- 3) When any points are not remains ,the 1st step is completed and one group is form.
- 4) Then new cluster is form from previous step

$$J(V) = \sum_{i=1}^c \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$$

' $\|x_i - v_j\|$ ' is the Euclidean distance between x_i and v_j .
' c_i ' is the number of data points in i^{th} cluster.
' c ' is the number of cluster centers.

IV. RESULTS DETECTED

Using Naïve Bayes the following classes are formed.

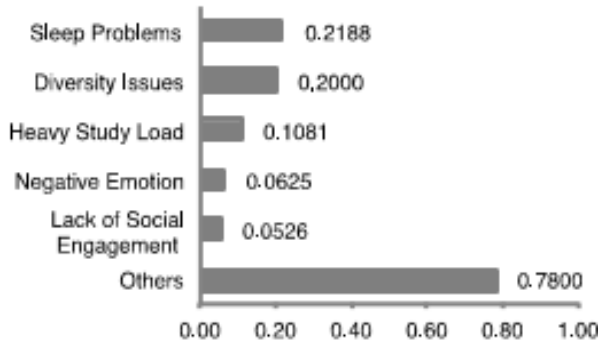


Fig 1 Classification result

Using K means the following clusters are formed.

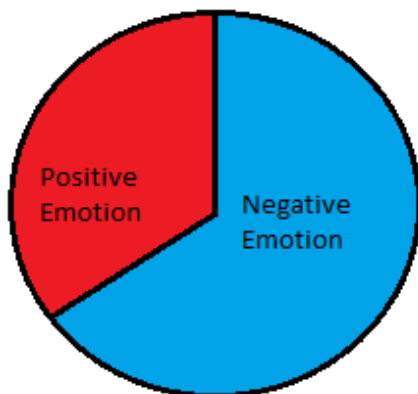


Fig2 Clustering result

V. CONCLUSION AND FUTURE WORK

Our study on the students post will help to inform the educational administrators to understand the experience of the student’s college experiences. This paper will provide the positive reviews as a feedback to the students and the negative reviews will be informed to the educational administrators. The future work could be analyzing the further student’s posts rather than the texts like images, gestures, videos etc.

REFERENCES

[1] Xin Chen, MihaelaVorvoreanu, and Krishna Madhavan.”Mning Social Media Data for Understanding Students ‘Learning Experiences” IEEE tamsactions on learning Technologies, ID, DOI 10.1109/TLT.2013.2296520.
[2] R. Baker and K. Yacef, “The State of Educational Data Mining in 2009: A Review and Future Visions,” J. Educational Data Mining, vol. 1, no. 1, pp. 3-17, 2009.
[3] G. Siemens and R.S. d Baker, “Learning Analytics and Educational Data Mining: Towards Communication and Collaboration,” Proc. Second Int’l Conf. Learning Analytics and Knowledge, pp. 252-254, 2012.

[4] Po-Wei Liang and Bi-Ru Dai.” Opinion mining on social Media data” 2013 IEEE 14th International Conference on Mobile Data Management.
[5] Ferguson, “The State of Learning Analytics in 2012: A Review and Future Challenges,” Technical Report KMI- 2012-01, Knowledge Media Inst. 2012.
[6] S. Cetintas, L. Si, H. Aagard, K. Bowen, and M. Cordova-Sanchez, “Micro-blogging in Classroom: Classifying Students’ Relevant and Irrelevant Questions in a Micro-blogging-Supported Classroom,” IEEE Trans. Learning Technologies, vol. 4, no. 4, pp. 292-300, Oct.- Dec. 2011.
[7] D. Gaffney, “#IranElection: Quantifying Online Activism,” Proc. Extending the Frontier of Society On-Line (WebSci10), 2010.