ISSN:2348-2079

Volume-6 Issue-2

International Journal of Intellectual Advancements and Research in Engineering Computations

# Analysis of Student Learning Experience by Mining Social Media Data

[1] P.Uma [2] Lavanya.V, [3] Evangelin Blessy.T, [4]Blessy.K

[1] Assistant Professor , [2-4] UG Students

Department of Computer Science and Engineering, Nandha Engineering College, Erode, India

## *Abstract*

**Many issues like depression, suicide, anger, are increasing among students. These issues are necessary to seek out and analyze, but students never discuss their issues with anyone. Today Social media is very popular medium where individuals share their feeling and opinion. Students also terribly active on social sites like Face book and Twitter. Their unceremonious discussion on social media (e.g. Twitter, Face book) illuminates light on their educational experiences—vote, sentiment, opinions, feelings, and concerns about the learning process. Data from such environments can supply valuable information which is helpful knowledge to understand student learning experiences. Analyzing such data can be challenging. The augmenting scale of data demands automatic data analysis techniques. This paper depicts a workflow to integrate both qualitative analysis and large-scale data mining techniques. This Paper emphasized on student's twitter posts to learn problems in student life as well as positive things occurred in their educational life. First conducted a qualitative analysis on sample tweets related to student's college life. Students face issues such as heavy work load of study, lack of social engagement, and sleep deprivation, employment issue, etc. In this paper "positive things" happen in student's life is also taken in to consideration. To classify tweets reflecting student's problem multilabel classification algorithms is implemented. Naïve Bayes and Linear Support Vector Machine Learning algorithms are used. The performance of these algorithms is compared in terms of accuracy, precision, recall and F1-Measure.**

**Support Vector Machine learning algorithm have more accuracy than Naïve Bayes Algorithm.**

***Index Terms***— Education, computers and education, social networking, web text analysis, Twitter Multi-Label Classifier.

## I. INTRODUCTION

Data mining research provides several techniques, tools, and algorithms for enormous amounts of data to answer real-world issues. Social media plays powerful role in today's era. As social media is generally used for various purposes, vast amounts of user created data can be made available for data mining. Main objectives of the data mining procedure are to communally handle large-scale data, extract useful patterns, and gain required knowledge. Social media sites such as Twitter, Face book, and YouTube provides stage to share happiness, struggle, sentiment, stress and acquire social support. On various social media sites, students discuss about their daily encounters in a comfortable and informal manner. They share their happiness and sorrows related to studies on social media in the form of judgmental comments, tweets, posts etc. Student's digital information provides large amount of implicit useful and reliable information for educational researchers to understand student's experiences outside the closed environment of classroom. This understanding can enhance education standard, and thus improve student employment, preservation, and accomplishment. The Massive quantity of information on social sites gives prospective to recognize student's problem, however conjointly promotes some methodological complexities in use of social media data for educational reasons. The complexities such as assortment of Internet slangs, absolute data volumes and moment of

students posting on the web. The research goal of this learning are:- a) To make the enormous amount of data useful for educational purpose, as well as to combine both qualitative analysis and large-scale data mining techniques. b) To examine student's informal tweets on twitter in order to investigate the problems and issues faced by students in their life.

## II. LITRATURE SURVEY

The theoretical foundation for worth of informal data on the web is drawn from Goffman's theory of social performance. Although developed to elucidate face-to-face interactions, Goffman's theory of social performance is widely used to explain mediated interactions on the web today. One amongst the foremost basic aspects of this theory is that notion of front-stage and back-stage of people's social performances. Compared with the front-stage, the relaxing atmosphere of backstage typically encourages more spontaneous actions. Whether a social setting is front-stage or back-stage could be a relative matter. For students, compared with formal classroom settings, social media is a relatively informal and relaxing. When students post content on social media sites, they usually post what they think and feel at that moment. In this sense, the data collected from on-line conversations may be more authentic and unfiltered than responses to formal research prompts. These conversations act as a zeitgeist for students' experiences. Many studies show that social media users may purposefully manage their on-line identity to "look better" than in real life. Human identity is complex and multifaceted. Humans acquire identity through social interaction and enact different roles, depending on context and social teams. For example, one may be a commanding, determined business executive at work, caring mother at home, and funny friend at a social gathering. 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. Other studies show that there's an absence of awareness regarding managing online identity among college students, and that young people typically regard social media as their personal space to hang out with peers outside the sight of parents and teachers. Students' online conversations reveal aspects of their experiences that are not simply seen in formal classroom settings, thus are usually not documented in academic literature. Researchers from diverse fields have analyzed twitter content to generate specific knowledge for their respective subject domains. For example, Gaffney analyzes tweets using histograms, user networks, as well as frequencies of top keywords to quantify online activism. Similar studies are conducted in different fields including healthcare, marketing, and athletics, just to name a few. Analysis methods used in these studies usually include qualitative content analysis, linguistic analysis, network analysis, and few simplistic methods such as word clouds and histograms. This model was then applied and validated on a brand new data set. Therefore, emphasize not only the insights gained from one data set, but also the application of the classification algorithm to other data sets for detecting student issues. The human effort is thus augmented with large-scale data analysis. Researchers have analyzed twitter data and place efforts for introducing methods for classifying twitter data For example, Alec Go introduced a completely unique approach for automatically classifying the sentiment of Twitter messages. These messages are classified as either positive or negative with relevance to a query term. This is useful for consumers who want to re- search the sentiment of products before purchase, or feedback to be aggregated without manual intervention.

## III. PROPOSED SYSTEM

This approach uses totally different machine learning classifiers and has feature extractors. The proposed system works on student's tweets. These tweets are related to student's educational experiences. This system defines seven labels which are: heavy study load, lack of social engagement, negative emotions, sleep problem, diversity issues, positive things and struggle. The objective is to explore student's informal conversations on twitter in order to understand issues and problems of students encounter in their learning experiences. The tweets are loaded and processed by standard text mining procedure called Pre-processing. This system is used to understand student learning experiences. The existing system uses classification algorithms finds only negative emotions of students learning, whereas the proposed system will use to discover positive as well as negative emotions. Naïve bayes and support machine algorithm is used to discover the positive as well as negative emotion of student about their learning experiences. The comparison of the results of these algorithms is done using parameters accuracy, precision, recall and F1- Measure. The figure 1 shows the architecture of proposed system. First step is to collect the data for processing. This data is nothing but student's tweets which have positive and negative expression about their academic experiences. In next step collected data is explored and define the categories into which tweets can be differentiated. The tweets are pre processed i.e. stemming, stop word cleaning. Stemming reduces inflected words to their stem, base or root form. In stop word cleaning, there are list of stop words which are removed by pre processing from text documents. However, on tokenization stream of text is break into words, phrases or symbols. The model is trained using multilabel classifier. The multi-label Support Vector Machine and Multi-

label Naive Bayes classifier are implemented and compared. The procedure of the proposed system is as follow:

1) In the step one data collection is done from twitter.

2) Inductive Content analysis procedure is performed and categories are identified.

3) The pre processing and is calculated in the step 2.

4) Naïve Bayes classifier, Linear SVM is applied on dataset in order to demonstrate its application in detecting student's issues is step 4.

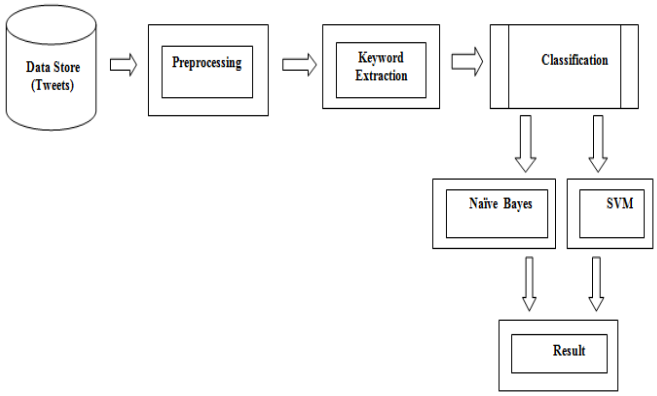


**Figure.1. System Architecture**

## IV. IMPLEMENTTATION

This project implements using following processes Inductive content analysis is used to identify the relevant and irrelevant tweets. In this categories are identified in which tweets are going to classify. Naïve bayes and support vector machine algorithm are used to classify the tweets of student's informal discussions.

### 4.1 DATA COLLECTION

It is challenging to collect social media data related to student's experiences because of the irregularity and diversity of the language used. Data searched on Twitter. The Twitter APIs can configure to accomplish this task. The search process was exploratory. Data searching is based on different Boolean combinations of possible keywords such as engineer, students, campus, class, homework, professor, and lab. Twitter API downloaded from twitter using tool My Twitter Scrapper. User should have twitter account and have to make an application for downloading tweets. These tweets can be download using hash tags. Twitter API can also be found location wise and particular person's account. My Twitter Scrapper is java based tool which is used to download the Twitter API. These tweets are saved in excel format. But this data is unclean and contain many errors as well as bugs. Similarly all tweets are not relevant to this system. System required tweets which contain positive as well as negative emotion or experiences of student in their educational life. Using My Tweet Scrapper Tool 25000 tweets are downloaded among these tweets only 1700 tweets are useful for this system.

### 4.1 Inductive content Analysis

Social media content like tweets contain a bulky amount of informal language, sarcasm, acronyms, and misspellings, meaning is often ambiguous and subject to human interpretation. Rostetal argue that in large scale social media data analysis, faulty assumptions are likely to arise if automatic algorithms are used without taking a qualitative look at the data. According to study there is no appropriate unsupervised algorithms could reveal in-depth meanings in our data. For example, LDA (Latent Dirichlet Allocation) is a popular topic modeling algorithm that can detect general topics from very large scale data. LDA has only produced meaningless word groups from our data with a lot of overlapping words across different topics. There were no pre-defined categories of the data, so it is necessary to explore what students were saying in the tweets. Thus, first step is to perform an inductive content analysis on the dataset. Inductive content analysis is one popular qualitative research method for manually analyzing text content.

### 4.2 CATEGORIES OF DATA

As a consequence proposed system first conducted an inductive content analysis on the n dataset. This paper classified the student tweets in to seven categories. Existing system have five prominent themes and proposed system consist of seven prominent categories:

#### i. Heavy Study Load

Analyses show that, classes, homework, exams, and labs control the student's life. Libraries, labs, and the college building are their most frequently visited places. Some illustrative tweets are "Study for 30 hours.., "Doing homework since morning still incomplete,", and "OS project due Thursday.", and "home work never finish". Students express a very stressful experience. Not being able to manage the heavy study load. This finding echoes a previous study on students' life balance by which indicates student's desire a more balanced life than their academic environment allows.

#### ii. Lack of Social Engagement

The analyses show that students have to give

up the time for social engagement in order to do homework, and to prepare for classes and study for exams. For example, "I feel like I'm hidden from the world—life of an student". Lack of social engagement is also tangled with the conventional nerdy and anti-social image of students. Some students embrace the anti-social image, while most others desire more social life as the examples above show.

**iii. Negative Emotion**

There are a bunch of negative emotions flowing in the tweets. This category specifically contains negative emotions such as hatred, anger, stress, sickness, depression, disappointment, and hopelessness. Students are mostly stressed with schoolwork. For example, "looking at my grades online makes me sick", "40 hours in the library in the past 3 days. I hate finals", and "I feel myself dying, #nervous". It is necessary for students to get help with how to manage stress and get emotional support.

**iv. Sleep Issues**

Analyses find that sleep issues are widely common among students. Students frequently suffer from lack of sleep and nightmares due to heavy study load and stress. For example, "Napping in the common room because I know I won't sleep for the next three days", "If I don't schedule in sleep time, it doesn't happen", and "I wake up from a nightmare where I didn't finish my physics lab on time". Chronic lack of sleep or low-quality sleep can result in many psychological and physical health issues; therefore this issue needs to be addressed.

**v. Diversity Issues**

The Analyses suggest students perceive a significant lack of females in engineering. For example, "eighty five kids leaving the classroom before mine. of those 85, four are girls. Engineers math class #Stereotypical", and "Keeping up with tradition: 2 girls in a class of 40". Male students in engineering are regarded as bad at talking with female students, because they usually do not have many female students around in their class. For example, "I'm sorry. We're not use to having girls around", "I pity the 1 girl in my lab with 25 guys. The issue here is not lack of diversity, but rather that students have difficulties embracing the diversity, because of many culture conflicts.

**vi. Positive**

A large number of tweets fall under this category. As student faces many issues in their educational life to recognizing as well as reducing those issues this system designed. Student also faces many positive things and emotions in their educational life tore cognize such emotions are equally important. By recognizing such emotions which can have positive impact on student life is to make students life happy. This category is consisting of emotions such as happy, Fine, good, easy, got, holiday, over, etc.

**vii. Struggle**

According to studied dataset conclusion occurs that many student tweets on their struggle in education. So new class introduced in proposed work. Student Tweets on they face struggle in finding job. They are struggling for passing exam. Syllabus is hard and problem in finding syllabus. For example: "Engineering is struggle in every field from finding syllabus to finding job, If I could stop time would catch up on my I tutorials so hard". This kind of tweets included in this category.

### 4.3 TEXT PREPROCESSING

Twitter users use some special symbols to convey certain meaning. For example, # is used to indicate a hash tag, @ is used to indicate a user account, and RT is used to indicate a re-tweet. Twitter users sometimes repeat letters in words so that to emphasize the words, for example, "huuungryyy", "sooo muuchh", and "Monnndayyy". Besides, common stop words such as "a, an, and, of, he, she, it", non-letter symbols, and punctuation also bring noise to the text. So we pre-processed the texts before training the classifier.

## V. EXPERIMENTAL RESULT

This system uses approximately 1700 tweet of student downloaded from twitter which contain student emotions. Firstly system preprocesses the data than support vector machine as well as naïve bayes algorithm applied on data. Performance analysis is done using k-fold cross validation. It is also called as rotation validation. In this case k has taken as 5. Hence 1700 tweets data is divided in to 5 folds. Table II shows the cross validation result of SVM. Here 1700 dataset is divided in to 5 fold. In first iteration 273 tweets are considered as test set and remaining 1370 are training set accuracy of this iteration is 90.35. In second iteration 273 tweets are considered as test set and remaining 1370 are training set accuracy of this iteration is 90.14.

**Table.1.cross validation result of naïve bayes.**

| FOLD | TEST SUBJECT | TRAIN SUBJECT | ACCURACY |
|---|---|---|---|
| 0 | 270 | 1373 | 83.18 |
| 1 | 273 | 1370 | 83.34 |
| 2 | 273 | 1370 | 83.18 |
| 3 | 273 | 1370 | 83.26 |
| 4 | 273 | 1370 | 83.12 |

Table I shows the cross validation result of Naïve Bayes Algorithm. Here 1700 dataset is divided in to 5 fold. In first iteration 270 tweets are considered as test set and remaining 1373 are training set accuracy of this iteration is 83.18. In second iteration 273 tweets are considered as test set and remaining 1370 are training set accuracy of this iteration is 83.34. In third iteration 273 tweets are considered as test set and remaining 1370 are training set accuracy of this iteration is 83.18. In forth iteration 273 tweets are considered as test set and remaining 1370 are training set accuracy of this iteration is 83.26. In fifth iteration 272 tweets are considered as test set and remaining 1370 are training set accuracy of this iteration is 83.12. In third iteration 273 tweets are considered as test set and remaining 1370 are training set accuracy of this iteration is 90.48. In forth iteration 273 tweets are considered as test set and remaining 1370 are training set accuracy of this iteration is 90.16. In fifth iteration 273 tweets are considered as test set and remaining 1370 are training set accuracy of this iteration is 90.35.

**Table.2. Cross validation result of svm.**

| **FOLD** | **TEST SUBJECT** | **TRAIN SUBJECT** | **ACCURACY** |
|---|---|---|---|
| 0 | 273 | 1370 | 90.35 |
| 1 | 273 | 1370 | 90.14 |
| 2 | 273 | 1370 | 90.48 |
| 3 | 273 | 1370 | 90.16 |
| 4 | 273 | 1370 | 90.35 |

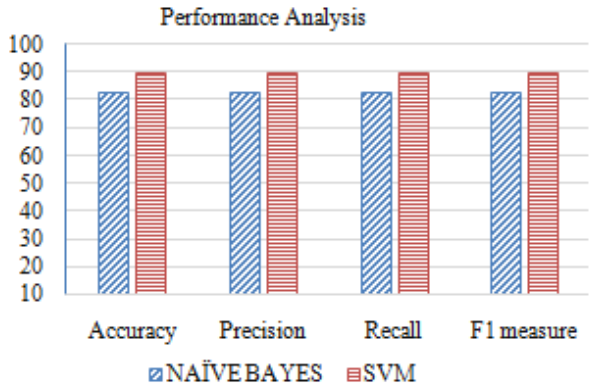


**Figure.2.Performance Evaluation of Nb and Svm**

## VI. CONCLUSION

This paper is classifies the student tweets to understand their problems as well as positive things in their educational life. Database is collection of tweets these tweets which depicts the information of student experiences in their educational life. The Tweets are stored in database. Preprocessing is done on tweets. Then classification techniques (Naïve Bayes and support vector Machine) are applied on data. In classification student tweets classified in to seven categories. That is negative, sleeping problems, positive, diversity issues, Lack of social awareness, heavy study load, Struggle. Cross validation technique is used to evaluate performance of a system. Accuracy of naïve bayes algorithm in five iteration is 83.18, 83.34, 83.18, 83.26, and 83.12 respectively. Accuracy of SVM in five iteration is 90.35, 90.14, 90.48, 90.16, and 90.35 respectively. Finally average of 5 iteration result is calculated to evaluate performance of a system. Accuracy, Precision, Recall And F1-Measure of Naïve Bayes is 83.12, 83.25, 83.04 and 83.21 respectively and the Accuracy, Precision, Recall And F1-Measure of SVM is 90.35, 89.94, 90.00 and 90.30 respectively Comparison of algorithm is done on the values of Accuracy, Precision, Recall, and F1 Measure and came to conclusion that Support Machine Algorithm gives more accurate prediction than Naïve Bayes Algorithm. Accuracy of Naïve Bayes is 83.12 % while Support Vector Machine is 90.35 %. Precision of Naïve Bayes is 83.25% while Support Vector Machine is 89.94 %. Recall of Naïve Bayes is 83.04 % while Support Vector Machine is 90.00 %. F1- Measure of Naïve Bayes is 83.21 % while Support Vector Machine is 90.30 %.

## REFERENCES

*BASIC FORMAT FOR BOOKS*

*:*

[1]. Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan, "Mining social media data for understanding students' learning experiences", IEEE Transaction, 2014.

[2]. Mariam Adedoyin-Olowe,, Mohamed Medhat Gaber and Frederic Stahl," A Survey of Data Mining Techniques for Social Network Analysis", School of Computing Science and Digital Media, Robert GordonUniversity Aberdeen, AB10 7QB, UK

[3]. E. Goffman, The Presentation of Self in Everyday Life. Lightning Source Inc., 1959.

[4]. J.M. DiMicco and D.R. Millen, "Identity Management: Multiple Presentations of Self in Facebook," Proc. the Int'l ACM Conf. Supporting Group Work, pp. 383-386, 2007.

[5]. M. Vorvoreanu and Q. Clark, "Managing Identity Across Social Networks," Proc. Poster Session at the ACM Conf. Computer Supported Cooperative Work, 2010.

[6]. B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short Text Classification in Twitter to Improve Information Filtering," Proc. 33rd Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 841-842, 2010.

[7]. Bodong Chen, Xin Chen, Wanli Xing, "Twitter Archeology of Learning Analytics and Knowledge conference on lerning analytic and knowledge, pp. 340-349,2015.

[12]. A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification Using Distant Supervision," CS224N Project Report, Stanford pp. 1-12, 2009.

[8]. R. Baker, "Educational Data Mining: An Advance for Intelligent Systems in Education" Teachers College, Columbia University

[9]. H. Loshbaugh, T. Hoeglund, R. Streveler, and K. Breaux, "Engineering School, Life Balance, and the Student Experience," Proc. ASEE Ann. Conf. & Exposition, 2006.

[10]. David M W Powers," Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation", Technical Report SIE-07-001 December 2007.

[11]. Ron Kohavi," A Study of CrossValidation and Bootstrap for Accuracy Estimation and Model Selection", appears in the International Joint Conference on Artificial Intelligence IJCAI.

[13]. Hsin-Ying Wu, Kuan-Liang Liu and Charles Trappey " The Theory On User Feedback Analysis."

[14]. Suleyman Cetintas, Luo Si, Hans Peter Aagard, Kyle Bowen, and Mariheida Cordova-Sanchez," Microblogging in a Classroom: Classifying Students' Relevant and Irrelevant Questions in a Microblogging-Supported Classroom," IEEE Transactions on Learning Technologies, Vol. 4, No. 4, October-December 2011

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

[16].W. Zhao, J. Jiang, J. Weng, J. He, E.P. Lim, H. Yan, and X. Li,"Comparing Twitter and Traditional Media Using Topic Models," Proc. 33rd European Conf. Advances in Information Retrieval, pp. 338- 349, 2011.

[17]. Andreas Hotho," A Brief Survey of Text Mining" KDE Group University of Kassel May 13, 2005.

.