*Turkish Journal of Computer and Mathematics Education Vol.12 No.12 (2021), 3110-3118*

*Research Article*

# Supervised Learning Techniques for Classification Of Students' Tweets

**Blessa Binolin Pepsi M[a], Senthil Kumar N[b]**

[a] Assistant Professor (Senior Grade), [b] Senior Professor
[a] Department of Information Technology, 2Department of Electrical and Electronics Engineering
[a,b] Mepco Schlenk Engineering College, Sivakasi, Tamilnadu.
[a] mblessa@mepcoeng.ac.in

**Article History:** Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 23 May 2021

**Abstract:** In today's era, up-to-date information can be retrieved from social network, internet community and data forums. People especially the younger generation share their feelings, happiness, experience and also day to day happenings in the social media platforms like Twitter. There exists large volume of unstructured data in it. The proposed system concentrates on the learning process of the engineering students and the problems faced by them during their study from their twitter posts. Since the data collected is huge, Apache hadoop map reduce environment is used for processing. The system includes pre-processing of tweets, calculating F1 measure, identifying prominent categories, identifying word and category probability and finally classifies tweets to the respective categories. The supervised learning techniques such as multiclass SVM based Platt Scaling, Naïve Bayes and logistic regression are used to identify heavy study load, lack of social engagement and sleep problems. Comparing the results attained, SVM achieves an accuracy score of 84% which is 5 to 10 percent higher than Logistic Regression and Naïve Bayesian method.

**Keywords:** Map reduce, Data Pre-processing, Large scale data Analytics, Classification, Support Vector Machine, Label Based Classification

## 1. Introduction

Students' digital footprints through various forums like Twitter, Facebook, and Blogs helps the educational researches to value the students' experience other than classroom surrounding. This may help the institution to comprehend and figure out the at-risk students thereby it can help to improve the quality of education, recruitment and retention of students. The massive social media data [1] paves a way to interpret students' experiences by making sense of data for educational purposes. Due to wide diversity of students across the world, manual analysis is impossible in the fast emergent scale of data, while conventional algorithms cannot capture it easily.

Traditionally, the schemes used for learning the students' experiences were through surveys, interviews and classroom goings-on. The scale of this data is inadequate since collected manually and its time consuming too. They cannot be observed with past data for analytics as it's collected in a particular time.

Existing learning analytics techniques focuses on evolving decision making process through course management systems and class technology usages [4]. The clear goal of this system is to infer and analyze the learning habits of students from a social networking environment [3] as it is an uncontrolled space to share their inner emotions and feelings at any point of time. The research goal of this is to infer their problems since students' are future workforce and their knowledge impacts on nation's economic growth in the years to come.

Twitter being in-style social media, is used to collect the data stream related to students experience. The hash tags [5] i.e. tag or topic starting with # sign is listed out and collected using APIs and each tweet is not more that 280 words hence concise too for analytics. Based on these, students can be helped to overcome the hurdles in learning

In this paper, large scale data analytics is done to analyze the tweets brought together using different classification techniques [6]. Classification is the process of organizing the data into categories to make it effective. Based on the experiences some categories are fixed initially to build the learning model in classification process. Next the tweets evolved from APIs related to students' problems are trained and tested to identify the category it belongs to. This workflow performs a qualitative analysis in order to identify at-risk student and to make proper decision for their future education by intervening their inferences.

The data retrieved can be huge and in turn it can be model of batch processing and stream processing depending on the application used. In this work, it is used as batch processing implemented on a map reduce framework in Apache Hadoop environment[10] The HDFS environment is used for data storage including namenodes, datanodes and secondary namenodes. For processing, the input data is splitted into map tasks and further does shuffling and sorting. The result of shuffle and sort is sent as reduce tasks to execute. The tasks work parallel in node managers. The job scheduling and resource allocation is managed by the resource manager inYARN. Finally the result of large scale data analytics is stored after map reduce task gets completed.

The paper is ordered as follows: the next section lists the various objectives of the proposed work. Section 3 describes the steps involved in the process Section 4 illustrates the analytical process identified to predict the model and Section 5 shows the application data used for implementation to detect at a specific university and comparison results of multi-label SVM classifier with Naïve Bayes and Logistic Regression. Section 6 concludes and lists the future work of the study.

## 2. Objective

The goal of this work is to classify the tweets based on the problems experienced by the students. The data is collected from forums like twitter and preprocessed and perform text classification into predefined categories using the steps followed in proposed system design.

The method must satisfy these requirements:

- To explore students learning experience through casual discussions made social media like facebook, twitter or any forums.
- To identify the main issues and troubles that students face during their study period
- To study and survey the needs for educational purposes in order to make something good from it.
- To perform qualitative analysis through large scale analytics techniques to attain knowledge.

Data is gathered from student's tweets and posts from the online forums. Since the forums are informal, students feel free to share their views in it. So, it's chosen for this proposed work to infer their experiences. Researchers across have taken the online social media content for analytics irrespective of subject domains to attain some knowledge from it. The analysis can be linguistic analysis, build word clouds or to analyse network traffic etc…

The proposed problem is based on supervised learning i.e. classification model to derive insight through content analysis. Any dataset can be used in the model to derive the students' intention based on large scale data analytics [7]. The classification algorithms taken in are Naïve Bayes, Logistic Regression and Support Vector Machine. Multi label classification is proposed with the methods above, since the tweet can fall into any defined category rather than binary classification [10]. From these labeled results, we can attain the students' social and economical views in all the aspects which in turn can help the society or the educational environment to guide and to improve in better ways.

## 3. Proposed System Design

The proposed includes different steps executed sequentially. They are listed below with the block diagram of it in Fig. 1

**Data Collection**- First data is collected from the twitter using #Engineering problems over a period of time.

**Data Storage** - Data is stored in some storage repositories.

**Qualitative Analysis** – Data is taken to perform inductive content analysis. The engineering student's problem was detected from the previous step. The prominent categories are chosen and tweets are classified

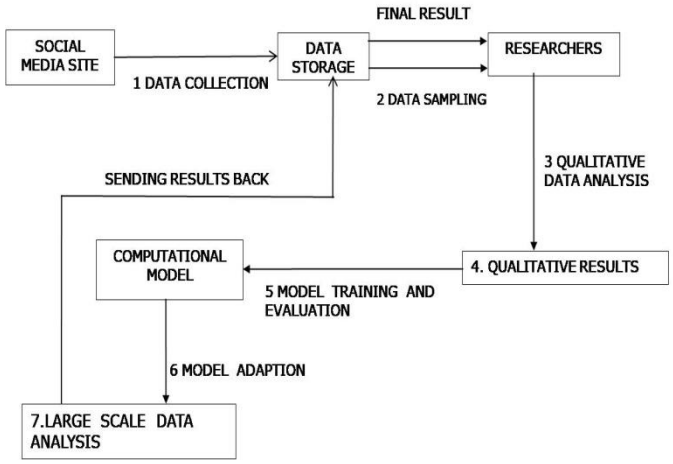



**Fig. 1:** Overall Workflow of Proposed System

**Model Training and Evaluation-** The performance of the classifier are calculated and evaluated using different supervised Learning Algorithms

**Model Adoption** – Uses supervised learning to train the model that assists in detection of engineering student problem.

**Large Scale Data Analysis Result -** The classification results are useful for researcher in decision making.

**Problem Definition**

The overall scope of the proposed work in the beginning is to acquire the informal social media data from all possible ways and then to preprocess them as structured data. Further the text mining is done by various steps and then analysed with different supervised learning techniques. The results infer the students' experiences to identify at-risk students and also to promote the students learning.

**Data Collection**

It is tough to collect from social sites due to mixture of languages used as a part of posts or tweets. The data collection was tried from Twitter using different hash tags like students, college, class, professor, laboratory work etc…Further a dataset retrieved using #engineering problems was identified which comprised the student experiences of Purdue University. There were 35,598 tweets of which some taken for training and other for testing using map reduce programming model

**Content Analysis**

Social media content is often ambiguous and Rost et al.[9] argues the same thing stating it is hard to perform analytics on that data. By research study, we could analyse that unsupervised learning algorithms finds different to infer right inner meaning of the data. Since unsupervised algorithms doesn't have any pre-defined categories to explore what feeling is expressed by the student. So, we decided to perform an inductive content analysis on the textual content i.e. retrieved tweets in order to define the student's mentality as different category. When categories were fixed, it was decided to check with the supervised learning technique for further analytics.

**Development of Categories**

The top themes used are: heavy study load, sleep problems, lack of social engagement, negative emotions and all other diversity issues like physical health problems, lack of motivation, future worries etc… For an example if the tweet, "lot of lab to complete" – is mentioned, this falls under the category of sleep problems and if the tweet is, "Why not Arts and Science?" – This falls under the category of negative emotion in case of engineering problems. This is how data is analyzed and classified based on the already existing learning model of classification algorithms

**Classification**

Supervised learning techniques works on text categorization i.e. given tweet is analyzed based on the content and assigned the set of predefined categories. It helps to get insights from the data and to automate the process of inferring students' views through informal tweets. The content analysis part requires measures like Kappa or F1 to perform multi-classification with the data which are not mutually exclusive. The techniques used for the process are Logistic Regression, Naïve Bayes' and Support Vector Machine.

## 4. Implementation

The proposed system is implemented with the following modules:

A. Data Preprocessing

B. F1 Measure

C. Frequent Word Collection

D. Evaluating Probability

E. Classification

F. Evaluation Measures

**Data Preprocessing**

First step is to preprocess the tweets in order to remove the noise like special symbols, http links and other informal conversations. The hash tags are removed from the tweets. In case of negative words it's removed and replaced with 'negtoken'. Further all links and symbols like @ are removed. The repeated letters are removed and

replaced with one letter if there is no meaning for it. A word 'Soooo' can be corrected as 'So' since the first word doesn't have a proper meaning. Stemming is done on the input data i.e. removal of 'ly', 'ing' etc… All the stops words are removed from the given input like 'a', 'or', 'the', 'an', etc… Once the data is preprocessed it proceeds for next analysis

**F1 Measure**

This measure is used for test accuracy. Here in this proposed model this is used to check whether a tweet can be added to a training set. The learning model of a classification problem is completely based on training set. The tweets included in training set should be accurate to perform analytics therefore a new logic is proposed to perform it using this accuracy F1 measure. After preprocessing, the words in the tweet are in use to calculate the F1 measure.

Consider two researchers have their own labels. The input (ie) number of labels from the researcher A for each tweet(x1) and input (ie) number of labels from the researcher B for each tweet(x2) is taken. Let the common labels between the two be (L) i.e. agreed labels between two.

Let p1 = L / x1 and p2 = L / x2

F1 measure is calculated by,

$$F1 = \frac{1}{N}\sum_{i=0}^{N}\frac{2p1.p2}{p1+p2}$$

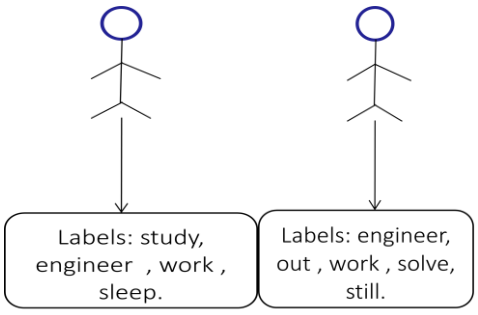
----- (1)

where N represents the number of tweets.

If the value of F1 is almost equal to 1 then the tweet can be accepted, else it can't be accepted to be added to the training set.

For example,

Given the labels for both researcher




Researcher 1 Researcher 2

Tweet - Hey BMEs engineer out hatred bigotry from human genome Thanks- everybody

Researcher A: engineer

Researcher B: engineer, out.

Common (L): engineer

p1 = 0.5 p2 = 1.0 f1 measure = 0.6667

Since the value of F1 measure is nearly 0.7, the tweet can be added to the training set for building the learning model.

**Frequent Word Collection**

The tweets that were chosen from the F1 measure score proceeds for the next step i.e. frequent word collection. The counts of each word in a tweet is taken and if the count is greater than the certain threshold value, then those frequently occurring words are alone separated. Those words are used to identify the prominent categories. This will define the category name for each tweet.

**Evaluating Probability**

The probability value of each category is calculated as follows, if there are M tweets out of which C tweets belong to a category c. then,

$$P(c) = \frac{C}{M} \qquad \text{------- ( 2 )}$$

Then the word probability is calculated in such a way,

If there are N words in a tweet, W={ $W_1$, $W_2$, … $W_n$} and L categories C={$C_1$,$C_2$, ..... $C_L$). If a word appears in category c for $m_{wnc}$ times, then pprobability of the word in specific category c is

$$p(w_n \mid c) = \frac{m_{w_n c}}{\sum_{n=1}^{N} m_{w_n c}} \qquad \text{--- (3)}$$

Similarly the probabilities of word in all categories are calculated.

**Classification**

Probabilistic supervised learning methods identify the most likely class the observation belongs to. Further, based on the above probability values evaluated if the probability value is greater than threshold value then the tweet will be mapped to the respective category. If two or more values are higher than threshold, then the one with higher probability is chosen. If no value is greater than the threshold value then the category 'diversity issues' will be mapped to it.

Applying this logic on Bayes' theorem,

Classify a document $d_i$ i.e. each tweet in testing set into category c or not c is verified using the following method

- The probability that $d_i$ belongs to category c is

$$p(c|d_i) = \frac{p(d_i|c).p(c)}{p(d_i)} \propto \prod_{k=1}^{K} p(w_{ik}|c).p(c). \qquad \text{--- (4)}$$

This process is repeated for all categories and the one which is larger than threshold T is accepted. If one or more categories are greater than T, then the one with higher probability is accepted. In case if nothing is larger than T, then a special category 'others' is used for it. This is how binary values can be interpreted for multi class classification

Applying this logic on logistic regression,

Uses logit sigmoid function, maps predictions to probabilities. The probability value lies between 0 and 1. Logistic regression has become a classification problem since a decision threshold is fixed based on the retrieved probability value.

The expected value on a hypothesis using the sigmoid function represents the logistic function as,

$$h(x_i) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \qquad \text{---- (5)}$$

We can define conditional probabilities for $d_i$ observation i.e. each tweet for every category is based on the following observation

$$P(y_i = 1 | x_i; \beta) = h(x_i)$$
$$P(y_i = 0 | x_i; \beta) = 1 - h(x_i) \quad \text{----- (6)}$$

Here y represents whether category is chosen or not and $x_i$ represents test sample

We can write it more compactly as:

$$P(y_i | x_i; \beta) = (h(x_i))^{y_I} (1 - h(x_i))^{1-y_i}$$

----- (7)

The probability of a tweet is calculated for all categories ($c_i$). Based on the above probability values evaluated if the probability value is greater than threshold value then the tweet will be mapped to the respective category. If two or more values are higher than threshold, then the one with higher probability is chosen. If no value is greater than the threshold value then the category 'diversity issues' will be mapped to it,

Applying the logic of Support Vector Machine i.e. a maximal margin hyperplane optimization technique,

Platt Scaling results in a probabilistic distribution over the classification models

It produces probability estimates

$$p(y = 1 | x) = \frac{1}{1 + \exp(Af(x) + B)}$$

------ (8)

i.e., a logistic transformation of the classifier scores $f(x)$, where $A$ and $B$ are two scalar parameters that are learned by the algorithm i.e. maximum likelihood estimation values evolved. Here it includes the values of categorical and word probability as vector.

Predictions are made for a category $C_i = 1$ iff $P(y=1|x) > ½$; if $B \neq 0$, the probability estimates is estimated using the normal values $C_i$ = sign ($f(x)$) where sign (f(x)) = Weight vector. (Word probability) + Bias.

Similarly if sign (f(x) > 0, then belongs to $C_i$ else doesn't below to that category.

Based on the above probability values evaluated if the probability value, the values above the marginal hyperplane is identified to be the right category. In case if two values are above, then the one higher is the best category.

**Evaluation Measures**

The performance of the classification models chosen for the proposed work includes accuracy, precision and recall. The overall prediction is based on confusion matrix,

| | Actual : Yes | Actual : No |
|---|---|---|
| Predicted :Yes | True positive(TP) | False positive(FP) |
| Predicted : No | False negative(FN) | True negative(TN) |

Here multilabel classification is done and measured using 2 ways, category based measure and tweet based measure.

In category based measure, for each tweet $d_i$ , the true set of categories falls under is Y and predicted set of categories from test set is Z.

$$Accuracy = \frac{1}{M} \sum_{i=1}^{M} \frac{Y_i \bigcap Z_i}{Y_i \bigcup Z_i}$$

--------- (9)

$$\text{Pr}\,ecision = \frac{1}{M}\sum_{i=1}^{M}\frac{Y_i \bigcap Z_i}{Z_i}$$

--------- (10)

$$\text{Re}\,call = \frac{1}{M}\sum_{i=1}^{M}\frac{Y_i \bigcap Z_i}{Y_i}$$

---------- (11)

In tweet based measure, each tweet on different categories were evaluated with the following measures i.e. overall accuracy of the model using all the supervised learning methods can be predicted

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

-------- (12)

$$precision = \frac{tp}{tp + fp}$$

-------- (13)

$$recall = \frac{tp}{tp + fn}$$

-------- (14)

## 5. Results and Discussion

The data collected from the experiences of Purdue University [12] students through Twitter API, 6500 tweets were taken for analysis and all the steps included in the proposed design was executed.

Any data related to students' experiences can be provided as an input to identify the prominent categories. The entire proposed system was implemented in Hadoop 2.6 version including Mapreduce programming under Ubuntu OS executed as single node cluster

Among the total tweets retrieved 72% was taken as training set and 28% was passed on test set i.e. nearly 1800 sample for testing. The evaluation measures accuracy, precision and recall were calculated in category based measure and tweet based measure.

The resultant category representation attained for the testing test is depicted below,

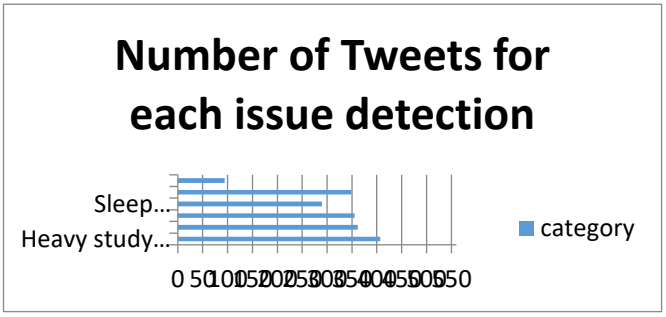


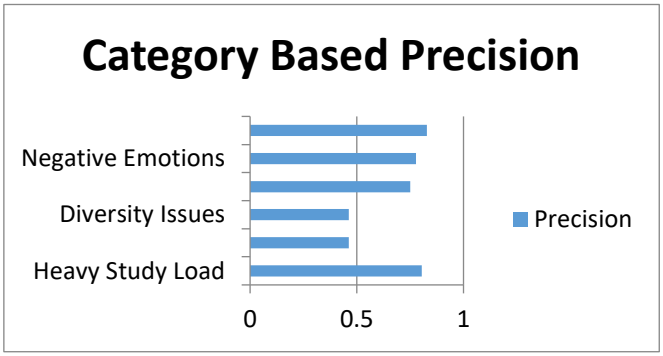

**Fig.2 :** Number of testing tweets in each category.

The results shows that most of the students have an inner feeling that study load is heavy for an engineering student. This prediction can help an educational researcher to recommend the institution to reduce the load and replace with many practical works. Being a digitized world, this proposed result of the student analysis based on tweets will play a vital role in future.

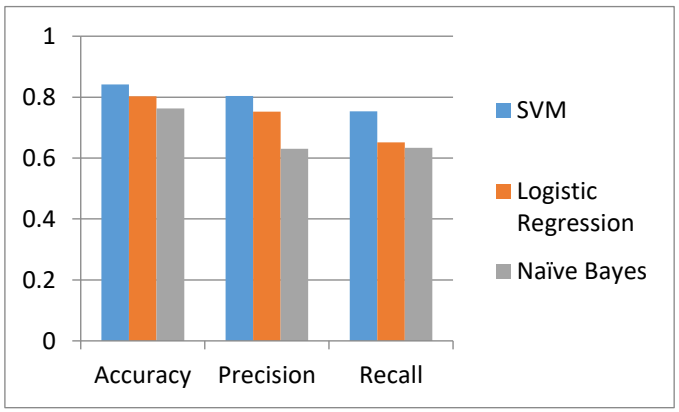
In category based measure, accuracy, precision and recall values of all categories are tabulated as table evaluated using SVM technique.

| **Category** | **Accuracy** | **Precision** | **Recall** |
|---|---|---|---|
| Heavy Study Load | 0.826 | 0.804 | 0.732 |
| Social Engagement | 0.842 | 0.463 | 0.95 |
| Diversity Issues | 0.944 | 0.462 | 0.857 |
| Sleeping Problems | 0.986 | 0.75 | 0.75 |
| Negative Emotions | 0.972 | 0.778 | 0.778 |
| Others | 0.854 | 0.829 | 0.684 |

The precision value calculated for each category is shown in the figure below. Higher the precision values for three prominent categories is shown below aprt from others.




In tweet based measure, the overall accuracy, precision and recall value calculated using the supervised learning techniques i.e. multiclass SVM, logistic regression and Naïve Bayes' classification is shown below.




The results depicts that among the probabilistic supervised learning techniques, support vector machine performs better than Logistic regression and Naïve Bayes classification technique.

## 6. Conclusion:

As we live in modern and digitized era, evolving an analytical output related to society's improvement paves way for the upliftment of future generation. The proposed system depicts the workflow to analyze the students'

experience through online social media forums. It lends a hand for the educational researchers and institutions to identify at-risk students added up can know the views of students' community. Prominent categories were identified and the tweets are classified using certain supervised learning techniques. Based on comparison Platt scaling multiclass support vector machine performs better than other techniques. The future of the work can include the process of semantic analysis that is not just counting keywords and classifying rather it can be classified by understanding the tweet. Most of the students upload memes or their experiences mostly as images in social media. Analysing the data through images can also be future work in upcoming days.

**References**

1. T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors," Proc. 19th Int'l Conf. World Wide Web, pp. 851-860, 2010.
2. Shihab Elbagir, And Jing Yang, "Twitter Sentiment Analysis Based on Ordinal Regression", IEEE Access, Volume 7, 2019, pp. 163677 – 163685.
3. 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.
4. Bo Liu, Keman Huang, Jianqiang Li, and MengChu Zhou, "An Incremental and Distributed Inference Method for Large-Scale Ontologies Based on MapReduce Paradigm", IEEE Transactions On Cybernetics, Vol. 45, No. 1, January 2015
5. Shenghua Liu, Xueqi Cheng, Fuxin Li, and Fangtao Li, "TASC:Topic-Adaptive Sentiment Classification on Dynamic Tweets", IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 6, June 2015, pp. 1696 – 1709.
6. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification Using Distant Supervision," CS224N Project Report, Stanford pp. 1-12, 2009.
7. M. Rost, L. Barkhuus, H. Cramer, and B. Brown, "Representation and Communication: Challenges in Interpreting Large Social Media Datasets," Proc. Conf. Computer Supported Cooperative Work,
8. pp. 357-362, 2013
9. L. T. Nguyen, P. Wu, W. Chan, W. Peng, and Y. Zhang, "Predicting collective sentiment dynamics from time-series social media," in Proc. 1st Int. Workshop Issues Sentiment Discovery Opinion Mining, 2012
10. S. Cetintas, L. Si, H. Aagard, K. Bowen, and M. Cordova-Sanchez, "Microblogging in Classroom: Classifying Students' Relevant and Irrelevant Questions in a Microblogging-Supported Classroom," IEEE Trans. Learning Technologies, vol. 4, no. 4, pp. 292-300, Oct.-Dec. 2011.
11. Ms.M.Blessa Binolin Pepsi, Ms.S.Haseena, Dr.S.Saroja, "Evolutionary Computation Access on Incremental Map Reduce for Mining Large Scale Data", International Journal of Recent Technology and Engineering (IJRTE), Vol.: 8, Issue No: 2S3; PageNo: 860-865

.