DECISION TREE BASED EVALUATION OF ENGINEERING STUDENTS LEARNING PROBLEMS FROM TWEETS

Vishnu K V[1] , Jishnu Kumar.A.P[2] ,Anjali P S[3]
vishnukv313@gmail.com,
Students, B.tech in Computer Science & Engg,
JBCMET Perumbavoor

Sujith M S
Assistant Professor,
Department of Computer Science & Engg,
JBCMET Perumbavoor

***Abstract*— Nowadays data mining plays important role in business and can also be used in professional educational purposes too. Informal conversations of engineering students including hash tags on social media like Twitter provides valuable information regarding their educational experiences, suggestions, feelings, and concerns about the learning process. Today the amount of data is increasing exponentially. So analyzing such data is very challenging.**

Our proposed work is done by twitter mining using a classification between our predefined data set and twitter data. A work flow is developed to integrate both qualitative analysis and large-scale data mining techniques. Engineering students also had a great impact upon these social medias. Heavy study load, lack of social engagement, and sleep deprivation are some of the problems found in engineering students. A Decision tree multi-label classification algorithm is being implemented to classify tweets reflecting the student's problems. This helps to study about these issues and how it is influenced upon social medias.

Keywords***: Twitter data mining, Engineering problem, Decision Tree***

I. INTRODUCTION

Through social media sites such as Twitter, Facebook students discuss and share their everyday experience in their own casual ways. Each Student’s digital activities will be different from others hence this’ digital footprints provide implicit knowledge and a whole new way of look towards education. These will provide to get the each student outside world experiences. This understandings can be useful to each professional colleges to recognize each of theirs student’s problems and can give necessary arrangements to improve their performance. The amount of information now available to crunch and parse in the service of analyzing absolutely anything is massive—and growing every second. So collecting and analyzing student details from social media will become more challenging. Traditional manual analysis cannot cope with this scale of data. So automatic analysis should be implemented.

Generally students’ performance is evaluated by conducting surveys, classroom activities and taking feedback from students. These are very time consuming and cannot do frequently and this type of activities can be conducted to a limited number of students. Hence these are not the ideal ways.

At this stage the introduction of data mining and advanced learning analytics can make efficient way of analyzing each student experience and come up with the problems that they are currently facing. The activities they are involving in social network like Twitter significantly related to their performance. Implementing these method institutions can provide the support for those who needs.

The research goals of this study are 1) to demonstrate a workflow of social media data sense-making for educational purposes, collect those data and classify them according to data set and 2) to explore engineering students’ informal conversations on Twitter, in order to understand issues and problems students encounter in their learning experiences on colleges.

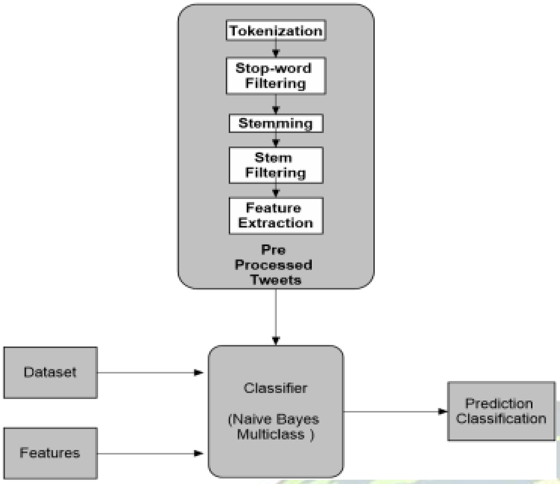

Fig. 1. The Work Flow of Existing System.

The existing techniques available are based on Naïve Bayes Multiclass Classifier. But the proposing system has more accuracy and performance. In this paper, By means of decision tree model, it has been found that the factors affecting each engineering student can be classified to a dataset of six categories. [6] discussed about Reconstruction of Objects with VSN. By this object reconstruction with feature distribution scheme, efficient processing has to be done on the images received from nodes to reconstruct the image and respond to user query. Object matching methods form the foundation of many state- of-the-art algorithms. Therefore, this feature distribution scheme can be directly applied to several state-of- the-art matching methods with little or no adaptation. The future challenge lies in mapping state-of-the-art matching and reconstruction methods to such a distributed framework. The reconstructed scenes can be converted into a video file format to be displayed as a video, when the user submits the query. This work can be brought into real time by implementing the code on the server side/mobile phone and communicate with several nodes to collect images/objects. This work can be tested in real time with user query results.

The remainder of this paper is organized as follows: the next section reviews theory of public discourse online, related work on text classification techniques used for analysing tweets, and data-driven approaches in education. Section 3 describes the proposed system (Fig. 2).The next section reviews theory of data collection, module description and section4 concludes this study.

II. RELATED WORK

A. *Mining Tweets*

People from different fields have been analyzing tweets in twitter for collecting information regarding specific subject domain. Brief review studies on Twitter from the fields of data mining, machine learning, and natural language processing are given below. They cover a wide range of topics including tweet classification [10], [11], [12], popularity prediction [13], [14], event detection [15], [16], topic discovery [17], [18], and [19] to name a few. Amongst these topics, tweet classification is most relevant to our study.

B. *Analystics Learning and Educational Data Mining*

Analytics learning and educational data mining (EDM) are latest data-based approaches emerging in education field. For institutions to make decision regarding to their students way of studying, these approaches can be used to analyse and collect the information from students educational environments. For example, researchers at Purdue University created a system named Signals that mines student performance data from Blackboard course system such as time spent reading course materials, time spent engaging in course discussion forums, and quiz grades [24]. Signals give students red, yellow, or green alerts on their progress in the course taken in order to promote self-awareness in learning. These are very time consuming approaches and they cannot be done frequently. Our proposed system extends the data scope of these data-driven approaches to include casual and informal social media data like tweets. We extend the understanding of students' experiences to the social and emotional aspects based on their informal online conversations. These are most important components of the learning experiences that are much less emphasized and understood compared with academic performance.

III. PROPOSED SYSTEM

The students learning system is an important part of most of the data mining techniques. The previous section dealt with various prior related works. After analysing those works, it is concluded that there is an immediate need to propose a new student learning system, namely, Decision tree based method for accurately analysing students learning problems from twitter data. The proposed method uses multi-class decision tree classification algorithm for analysing students learning problems. This algorithm is more efficient than multi-class Naive Bayes classification algorithm with respect to noise reduction. The proposed method uses Twitter API to retrieve tweets related to students learning problems. Initially, the retrieved tweets related to students learning related problems are saved into a csv file. Pre-processing stage removes the tweets does not specify students learning related problems, Re-tweets, noises, etc. After pre-processing extract important features using feature extraction methods such as count Vectorizer and Tf-Idf Transformer. Select top features from extracted features using Chi square feature selection methods. The selected features and labels are given to the Multi-class decision tree classifier for identifying students learning related problems. The detailed description of the proposed method is described in the following sections.

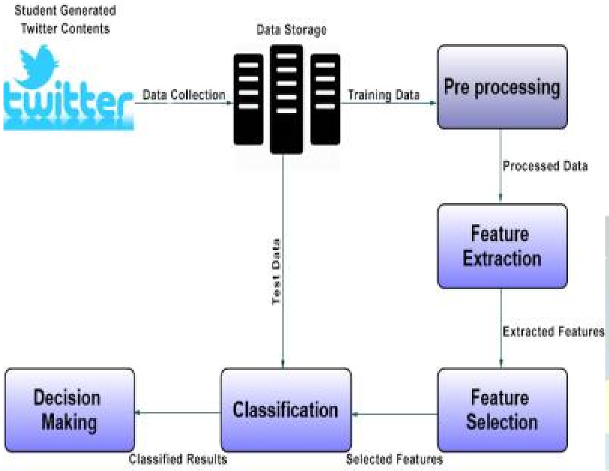

Fig. 2. The Work Flow Of Proposed System.

A. *Mining Tweets*

Manually classified the training dataset and identified major issues or problems which engineering students encountered in their college life. Major problems faced by engineering students fall into five prominent themes
They are: a) heavy study load, b) lack of social engagement, c) negative emotion, d) sleep problems, and e) diversity issues. Others indicates some problems other than five prominent classes.

1) Heavy Study load: Students are not able to handle the stressful life as it leads to lack of social engagement, lack of sleep, stress, depression, and some health problems.

2) Lack of Social Engagement: Social engagement in students helps the students in releasing the stress and therefore the students must be involved in doing the social works. Lack of the social works can result various problems among students which in turn result in anti-social image of engineers.

3) Negative Emotion: A negative emotion is being categorized only when emotions such as hatred, anger, stress, sickness, depression, disappointment, and despair were identified.

4) Sleep Problems: Lack of sleep which is widely seen in engineering students, results in many psychological and physical health problems. This arises because of heavy study load and stress.

5) Diversity Issues: One of the diversity issues includes the problem of understanding the lectures of foreign professors in the class.

6) Others: This indicates some other problems which are seen in engineering students in some cases which include curriculum problems, lack of motivation, procrastination, career and future worries, identity crisis, thought of switching majors, and physical health problems.

IV. MODULE DESCRIPTION

A. *Preprocessing*

Preprocessing steps helps in noise reduction in the dataset. Many symbols are being used by Twitter to convey special meaning [35]. For example, # is used to indicate a hashtag, is used to indicate a user account, and RT is used to indicate a re-tweet. Stop words ”a, an, and,of, he, she, it” non letter symbols, and punctuation also bring noise to the text. Thus the text are preprocessed before training the classifier: Detailed explanations of preprocessing are given below.

1) Replace all the emotions with their sentiment polarity.

2) Non-letter symbols and punctuation contained words are removed which includes the removal of @ and http links, white spaces and RTs.

3) In the hashtags only # sign is removed and the hashtag texts were kept as such.

4) For repeating letters in words, If we detected more than two identical letters repeating, we replaced them with one letter. Therefore, ”huuungryyy” and ”sooo” were corrected to ”hungry” and ”so”. ”muuchh”was kept as ”muuchh”. Originally correct words such as ”too” and ”sleep” were kept as they were.

5) Duplicating or repeating words were removed.

6) The common stop words were removed by using snow ball stemmer which includes the removal of ”the”, ”at”, ”never” etc.

B. Feature Extraction

At first, the processed tweets are analysed. In the feature extraction stage there occurs 2 stages, Count vectorizer and Tf-Idf Transformer. In the Count Vectorizer, features are sorted by name. From among these sorted list, the rare and common features are discarded. Now calculate the document frequencies for the features and finally create a feature vector. In the second stage, Tf-Idf \transformer, the features are transformed into tf representation. After that it is transformed into idf representation. Finally the Tf-Idf is calculated from Tf and Idf and thereby the features are extracted.

1) Count Vectorizer

CountVectorizer aims to Extract the vocabulary from a given collection of documents and generates a vector of token counts for each document.When an a-priori dictionary is not available, CountVectorizer can be used as an Estimator to extract the vocabulary and generate a CountVectorizer.

A CountVectorizer feature extractor assigns each occurring word in a corpus a unique identifier [37]. With this mapping, it can vectorize models such as bag of words or n-grams in a efficient way. The unique identifier assigned to a word acts as the index of a vector. The number of word occurrences is represented as a vector value at a specific index.

An n-gram model is a kind of probabilistic language model for predicting the next item in the order of an (n-1)-order Markov model. In the area of probability and computational linguistics, an n-gram is an adjacent order of n items from a given order of speech or text. The n-grams commonly are collected from a speech or text corpus. At the point when the items are words, n-grams may also referred to as shingles. An n-gram of size 1 is referred to as a unigram, size 2 is a bigram, size 3 is a trigram and so on.

2) TfidfTransformer

Tf implies term-frequency while tf-idf implies term frequency times inverse document-frequency. This is a typical term weighting scheme in information retrieval, that has additionally discovered great use in document classification. Transform a count matrix to a normalized tf or tf-idf representation. The objective of utilizing tf-idf rather than the raw frequencies of a token in a given document is scale down the effect of tokens that happen very frequently in a given corpus and that are henceforth experimentally less informative than components that happen in a little part of the training corpus.

A high weight in tf-idf is resulted by a high term frequency and a low document frequency of the term in the entire collection of documents; the weights subsequently tend to filter out common terms. Since the proportion inside the idfs log function is always greater than or equal to 1, the value of idf (and tf-idf) is greater than or equal to 0.

C. Feature Selection

Feature selection criteria is done with the help of Chi-square method. At first the chi-squared score of the extracted features are calculated and among them, K best features are selected and it is finalized as the selected features.

Chi-square: Compute chi-squared stats between each non-negative feature and class. This score can be used to select the n features with the highest values for the test chi-squared statistic from X, which must contain only non-negative features such as booleans or frequencies (e.g., term counts in document classification), relative to the classes.

D. Classification

In decision tree multi class classifier, selected features are considered and it is provided towards the classification purpose. From the selected features, information is calculated. After that it calculates the entropy and gain. From these results, it classifies the category. One-versus-all or Binary Relevance is one of the transformation methods which consists of assuming the independence among categories, and train a binary classifier for each category. All kinds of binary classifier can be transformed to multi class classifier using the one-versus-all heuristic. table footnotes.

In decision tree algorithm Shannons information theory was also included and it gains the higher performance too. Selection of the attributes from all levels of decision tree was done by using information gain criteria. Attribute with the largest information gain is selected to make decision tree root nodes. The different values of the node are used for establishing branches. Thus according to the instances of various branches the decision tree nodes and branches are recursively built, until a certain subset of the instances belongs to the same category. Information based method depends on two assumptions. Let C contain p objects of class P and n of class N.

The assumptions are:

1) Any correct decision tree for C will classify objects in the same proportion as their representation in C. An arbitrary object will be determined to belong to class P with probability p/(p + n) and to class N with probability n/(p + n).
2) When a decision tree is used to classify an object, it returns a class. A decision tree can thus be regarded as a source of a message 'P' or 'N', with the expected information needed to generate this message given by

$$I(p,n) = -\frac{p}{p+n}log_2\frac{p}{p+n} - \frac{n}{p+n}log_2\frac{n}{p+n}$$

If attribute A with values [A 1 , A 2 ....A v] is used for the root of the decision tree, it will partition C into [C 1 , C 2 ....C v] where C i contains those objects in C that have value A i of A. Let C i contain p i objects of class P and n i of class N. The expected information required for the sub tree for C i is I(p i , n i). The expected information required for the tree with A as root is then obtained as the weighted average

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

where the weight for the i th branch is the proportion of the objects in C that belong to C i . The information gained by branching on A is therefore

$$gain(A) = I(p,n) - E(A)$$

A good rule of thumb would seem to be to choose that attribute to branch on which gains the most information. Decision tree examines all candidate attributes and chooses A to maximize gain(A), forms the tree as above, and then uses the same process recursively to form decision trees for the residual subsets C 1 , C 2 ....C v . A special case arises if C contains no objects with some particular value A j of A, giving an empty C j . Decision tree labels such a leaf as ”null” so that it fails to classify any object arriving at that leaf.A better solution would generalize from the set C from which C j came, and assign this leaf the more frequent class in C.

A straight forward method of assessing this predictive accuracy is to use only part of the given set of objects as a training set, and to check the resulting decision tree on the remainder. At each non-leaf node of the decision tree, the gain of each untested attribute A must be determined. This gain in turn depends on the values p i and n i for each value A i of A, so every object in C must be examined to determine its class and its value of A. Decision trees total computational requirement per iteration is thus proportional to the product of the size of the training set, the number of attributes and the number of non-leaf nodes in the decision tree. The relationship appears to extend the entire induction process, even when several interactions are performed. No exponential growth in time or space has been observed as the dimension of the induction task increase, so the technique can be applied to large tasks.

V. CONCLUSION

There are many limitations for the manual qualitative analysis and large scale computational analysis of user generated textual content. Machine learning based classifiers help the researchers in learning analytics, educational data mining, and learning technologies effectively. Social media data provides substantial details regarding learning problems relating to engineering students. This data can be extracted and analysed using machine learning classifiers. This technique can also be employed for identifying issues related to educational fields which can throw light to educational administrators, practitioners and other relevant decision makers to gain further understanding of engineering students' learning related experiences.

References

[1] Chen, Xin, Mihaela Vorvoreanu, and Krishna Madhavan. "Mining social media data for understanding students' learning experiences." IEEE Transactions on Learning Technologies 7.3 (2014): 246-259.

[2] X. Chen, M. Vorvoreanu, and K. Madhavan, "A Web-Based Tool for Collaborative Social Media Data Analysis," presented at the Third International Conference on Social Computing and Its Applications, Karlsruhe, Germany, 2013.

[3] M. Clark, S. Sheppard, C. Atman, L. Fleming, R. Miller, R. Ste-vens, 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.

[4] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education," *Educause Review*, vol. 46, no. 5, pp. 30– 32, 2011.K. Elissa, "Title of paper if known," unpublished.

[5] M. Rost, L. Barkhuus, H. Cramer, and B. Brown, "Representa-tion and communication: challenges in interpreting large social media datasets," in *Proceedings of the 2013 conference on Comput-er supported cooperative work*, 2013, pp. 357–362.

[6] Christo Ananth, M.Priscilla, B.Nandhini, S.Manju, S.Shafiqa Shalaysha, "Reconstruction of Objects with VSN", International Journal of Advanced Research in Biology, Ecology, Science and Technology (IJARBEST), Vol. 1, Issue 1, April 2015, pp:17-20

[7] C. J. Atman, S. D. Sheppard, J. Turns, R. S. Adams, L. Fleming, R. Stevens, R. A. Streveler, K. Smith, R. Miller, L. Leifer, K. Ya-

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

[9] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 40, no. 6, pp. 601–618, 2010.

[10] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 2010, pp. 241–249.

[11] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, pp. 1–12, 2009.

[12] K. Nishida, R. Banno, K. Fujimura, and T. Hoshide, "Tweet classification by data compression," in *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, New York, NY, USA, 2011, pp. 29–34.

[13] R. Bandari, S. Asur, and B. A. Huberman, “The pulse of news
in social media: Forecasting popularity,” presented at *The International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.

[14] L. Hong, O. Dan, and B. D. Davison, “Predicting popular messages in twitter,” in *Proceedings of the 20th international conference companion on World wide web*, 2011, pp. 57–58.

[15] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes Twitter users: real-time event detection by social sensors,” in *Proceedings of the 19th international conference on World wide web*, New York, NY, USA, 2010, pp. 851–860.

[16] H. Becker, M. Naaman, and L. Gravano, “Selecting quality Twitter content for events,” in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM’11)*, 2011.

[17] W. Zhao, J. Jiang, J. Weng, J. He, E. P. Lim, H. Yan, and X. Li, “Comparing Twitter and traditional media using topic models,” *Advances in Information Retrieval*, pp. 338–349, 2011.

[18] L. Hong and B. D. Davison, “Empirical study of topic modeling in Twitter,” in *Proceedings of the First Workshop on Social Media Analytics*, 2010, pp. 80–88.

[19] E. Pearson, “All the World Wide Web’s a Stage: The performance of identity in online social networks,” *First Monday*, vol. 14, no. 3, pp. 1–7, 2009.

[20] J. M. DiMicco and D. R. Millen, “Identity management: multiple presentations of self in facebook,” in *Proceedings of the 2007 international ACM conference on Supporting group work*, 2007, pp. 383–386.

[21] M. Ito, H. Horst, M. Bittanti, danah boyd, B. Herr-Stephenson, P. G. Lange, S. Baumer, R. Cody, D. Mahendran, K. Martinez, D. Perkel, C. Sims, and L. Tripp, “Living and Learning with New Media: Summary of Findings from the Digital Youth Project,” The John D. and Catherine T. MacAuthur Foundation, Nov. 2008.

[22] D. Gaffney, “#iranElection: Quantifying Online Activism,” in *WebSci10: Extending the Frontier of Society On-Line*, Raleigh, NC, 2010.

[23] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques,” in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, Stroudsburg, PA, USA, 2002, pp. 79–86.

[24] K. E. Arnold and M. D. Pistilli, “Course signals at Purdue: Using learning analytics to increase student success,” in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 2012, pp. 267–270.

[25] Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, “Short text classification in twitter to improve information filtering,” in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2010, pp. 841–842.