Analyse and Classify Rural Area Engineering Students Twitter Posts using Data Mining Technique

Mr. Sambhaji .D. Rane
Assistant Professor
Department of Information Technology
DKTE Society's Textile and Engineering Institute, Ichalkaranji Maharashtra, India

Abstract— Students' informal conversations on social media such as Twitter and Whatsapp, are useful for understand their learning experiences, and feelings. Data from such social media environments can provide valuable information about students learning system. Collecting and analyzing data from such media can be difficult task. However, the large scale of data required to automatic data analysis techniques for classify twitter data. Proposed new system is to combination of qualitative analysis and large-scale data mining techniques. This system focuses on engineering students' Twitter posts which are collected from rural area engineering colleges to understand issues and problems in their learning. First conduct a qualitative analysis using ML studio on tweets collected from engineering colleges using term #DStudents problems, engineering Problem, Aluminisuggestions and lady Engineer. Collected tweets are related to engineering students' college life. In proposed system used a multi-label classification algorithm to classify tweets reflecting students' problems such as soft skill issues, heavy study load, lack of social engagement, and sleep problems.

Keywords: Data Mining, Social Networking, Machine Learning, Tweet Analysis, Classification

I. Introduction

Social media site Twitter has become important venue for the young generation to communicate and exchange information about their learning process. On various social media sites, such as Twitter, Whatsapp students discuss and share their everyday problems in an informal manner. Students' written tweets provide implicit knowledge and a whole new view for educational analysts to understand students' learning experiences outside the controlled classroom environment. This understanding can inform institutional decision-making on interventions for at-risk students, improvement of education quality in college and thus enhance student recruitment, retention ratio and placement in companies [1].

The learning analytics and educational data mining fields have focused on data obtained from classroom technology usage, or controlled online learning environments to inform educational decision-making. Traditionally, educational analysts use methods to collect data for analysis such as surveys using offline forms, interviews with students, conduct various classroom activities to collect data related to students' learning experiences [2]. These methods usually require more time. The scale of such studies is usually limited. Twitter is a well liked social media site. Its content is mostly public and very short. Twitter provides free APIs for data stream. Therefore, in proposed system used Twitter for chose to start from analyzing students' posts.

Here more focus is given on engineering students' Twitter posts to find the problems in their educational experiences. Engineering colleges and departments have been struggling with students' recruitment, retention and placement issues. Based on understanding of issues and problems in students' life, policymakers can make decisions on services that can help students to overcome such problems and issues. Many issues such as study problems, lack of social engagement and soft skill issues clearly come out.

II. Proposed System

A proposed system is focuses on engineering students' Twitter posts to understand issues and problems in their educational experiences. The proposed scheme is made up of Twitter data extraction, tweets data cleaning. Classification of tweet data and web module .The proposed scheme performs various operations on tweets as shown in Fig.1

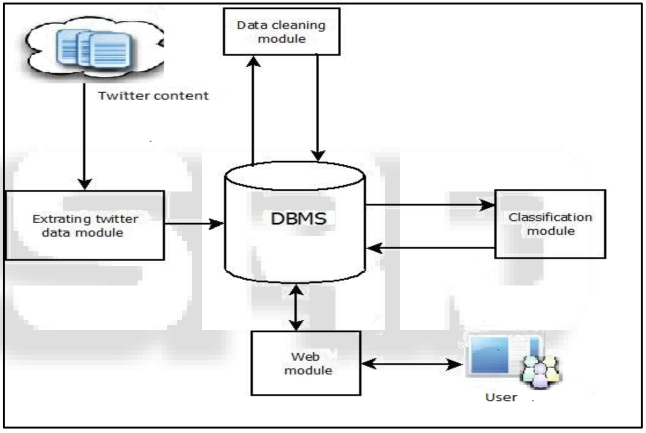

Fig. 1: Architecture of Proposed System for Mining Twitter Data.

In the first phase user extract tweets from twitter using twitter standard API[14]. Tweet processing operation performed in second phase. Then, tweet classification is perform using Naïve Bayes algorithm, tweets are classified into heavy study load, lack of social engagement, negative emotions, sleep problems, soft-skill issues and other. In data cleaning phase perform various operation on tweet to remove noise from it.

A. *Extracting Twitter Data*

Tweets searching started using possible terms such as engineer, students, college, class, homework, but the data set still contained more noise, also collected very small relevant tweets count. Found that, more relevant Twitter tweets are extracted using input query terms such as engineering Problem, DStudents problems, Alumini suggestions and lady Engineer with use of Twitter. The authentication of API requests on Twitter is carried out using OAuth. Detailed steps for making an API call from a Twitter application using OAuth is given below.

1) Applications are required to register themselves with twitter. Through this process the application is issued a consumer key and secret.

2) The application uses the consumer key and secret to create a unique Twitter link to which a user is directed for authentication. Twitter verifies the user's identity and issues an OAuth verifier.
3) The application uses the PIN to request an Access Token and Access Secret unique to the user.
4) Using the Access Token and Access Secret, the application authenticates the user on Twitter and issues API call.

The tweets extraction algorithm is given in Algorithm 1. In this algorithm used query term such as engineering Problem, DStudents problems, Alumini suggestions and lady Engineer as input for extract tweets from Twitter. The algorithm repeats itself recursively until there are no more tweets to be discovered. As a result, the output of algorithm tweets extraction is a set of tweets. These tweets are stored into database for further processing.

```
Algorithm TweetExtraction (term)
Input:  query term
Output: Tweets
// Extract tweets from the Twitter. Return true if successful.
begin
    // Build connection with Twitter for tweets Extraction.
    twitter: = TwitterFactory(cb.build()).getInstance()
    while result for query not null   do
        result := twitter.search(query)
        tweets := result.getTweets() return true
        for each tweet do
            insert tweets in to database
        end for
    end while
end
```

Algorithm 1: Tweets extraction.

B. Tweet Data Cleaning and Text Pre-processing

In this module pre-processed the texts and find useful text before training the classifier. Take input as tweets which are collect using Extracting Twitter Data module. Perform cleaning operation on tweets because there is noise present in to collected tweets so there to Pre-processing the tweets before training the classifier.

Use MySQL5.5 database for store collected and processed tweets. First database "social_data_mine" is created using create database social_data_mine command in sql. Under the "social_data_mine" database create various tables which are required for store data. Tables are tweet, search_topic and processing_tweet .Tweet table is used to store the collected tweets. Un-processed tweets are store in to tweet table. Search_topic table is used to store search topic name (query term) and topic id and processing_tweet table is used to store processed tweet.

Pre-processed the texts includes removed all the #engineeringProblems, #DStudentsproblems, #Aluminisuggestions and #ladyEngineer hashtags. For other co-occurring hashtags, only removed the # sign, and kept the hashtag texts. Remove all words from the tweets that contain non-letter symbols and punctuation. Tweets preprocessing done with help of remove special characters, stop words and stemming. After removing unnecessary symbols from collected tweet all the data is inserted in to processing_tweet table for further processing.

1) Remove Http Link

There is no use of http link for tweet classification, so remove the http link from the collected tweets. Perform splitting operation on processed tweet message. After splitting scan the every word from processed tweet. If the word not starts with http then add into the result and return the tweet without http link.

2) Remove Special Characters

Remove all the #engineeringProblems hashtags. For other co-occurring hashtags, Only removed the # sign, and kept the hashtag texts, Removed all words that contain non-letter [0-9] symbols and punctuation, such as # ,:$,% ,?,/,>,=,!,|,(etc. Scan the processed tweet-I message, if the any special character found in the tweet message then it replace with the blank and return the tweet without special character.

3) Stemming

Stemming means reducing a word to its base (or stem).Stemming is useful when doing any kind of text analysis concerned about the content of a the different times of verbs and the different ending for singular and plural, make it difficult to discern the importance of specific words with in text when treat each word as it is .Use a dictionary that lists all words together with their stems. Wordnet dictionary is large lexical database of English Nouns, Verb, Adjectives and Adverbs. It is free and publicly available for download. For actually stemming dictionary and a morphological processor is used. If it returns null then word not present in wordnet dictionary. Find the word match do with help of LookupBaseForm().Get the wordstem form index word, index word used for organize the word in wordnet dictionary.

4) Remove stopwords

Remove the words that are very commonly used in a given language because to focus on the importance words. Tweet text contain stop words such as hi, etc, be, as and many more words. Keep the words such as all, always, much, more because tweet used these words frequently. Performing splitting operation on the processed tweet message, split tweet message in to separate word and scan the every word , if a word is not found in stopword list then it add into the result and return the tweet without stopwords. After processing all tweets, it store into processing_tweet table.

C. Tweets Classification

1) Categories Development:

There were no pre-defined categories of the data so need to explore what students were saying in the tweets. It is very challenging task to develop categories for classification model. For categories development, online Azure Machine Learning (AML) studio [15] is used. Model for find the topics from tweets is built using AML. Following processes is followed for topic modelling.

1) Create project for topic modelling
2) Upload the tweet data set
3) Choose the text analytics algorithm LDA for topic modelling

2) Latent Dirichlet Allocation (LDA):

LDA represent tweets as mixture of topic that spit out words with probability.LDA algorithm has automatically detected the topics that tweets contain. It is used for topic modeling.

LDA algorithm take input as topic count and n-gram term for classification tweets into topics number.

For topic name and qualitative result perform content analysis on topic result generated by text analytics algorithm such as Latent Dirichlet Allocation (LDA). Content analysis is to identify what are the major worries concerns and issues that engineering student encounter in their study and life. Prominent categories are identified after analysing result of AML studio tool are heavy study load, lack of social engagement, negative emotions, sleep problems, soft-skill issues and other.

Found that many tweets could belong to more than one category. For example, "Why I am not in art school? Hate being in engineering school. Too many stuff, No enjoyment" falls in to heavy study load, negative emotions at the same time. For that required to use multi-label classification.

The prominent categories are

a) Heavy Study load

Analysis shows that classes, homework, exams and lab dominate the student life. Libraries lab and the engineering building are most frequently visited places. Some illustrative tweets are "Study over 20 hours for test-I", "so much homework, so little time".

b) Lack of social engagement

Analysis shows that students need to sacrifice the time for social engagement in order to home work and to prepare for classes and examination. For example, "I feel like I am hidden from the world-life of engineering student".

c) Negative emotions

Only categories tweets as negative emotion when it specially express negative emotions such as hatred, anger, stress, sickness, depression, disappointment and despair. Students are mostly stressed with schoolwork. For example, "is it bad that before I started studying for my test today that i considered throwing myself in front of a moving car?"

d) Sleep problems

Sleep problems are widely common among engineering students. Student frequently suffer from lack of sleep and nightmares due to heavy study load and stress. For example, "I wake up from nightmares where I didn't finish my physical lab on time".

e) Soft-skill issues

Soft skill issues are widely common among rural engineering students. Student frequently suffer lack of confidence, lack of communication skill. For example, "required more training to improve communication skill".

f) Other

Tweet from this categories are do not have a clear meaning, do reflect various issues that engineering student have, but very small volume. For finding topics from this category AML studio is used. LDA algorithm is applied only on other category tweets, then gets tweets category related to curriculum problems, physical health problems, lack of gender diversity, lack of motivation, nerdy culture, identity crisis but count of tweets related to this categories are very small so, treated all this category as other. Tweet other than above five categories falls in to other category.

3) Tweets Classification

This model is used to classified tweets based on categories. Used multi label Naïve Bayes classifier to classified tweets based on categories. Take input as processed tweets. Apply classification algorithm on processed tweets for categories wise classification of tweets. Tweets are classifieds into prominent categories such as heavy study load, lack of social engagement, negative emotions, sleep problems, soft-skill issues and other. Algorithm result is stored into naivebayes table.

4) Naïve bayes (NB) multi-label classifier

Naïve Bayes is easy and fast to predict class of test dataset. It also performs well in multiclass prediction. When assumption of independence holds NB classifier perform better to other models [7]. It performs well in case of categorical input. Naïve Bayes classifier used for tweets classification. Basic procedure for multi-label classifier follows. Each tweet is considered as document, there are a total number of N words in the learning dataset tweets collection $W=\{w_{1,} w_2,\ldots,w_N\}$ and total number of L categories $C=\{c_1,c_2,\ldots\ldots c_L\}$.

Suppose there total number of M tweets in the training set and A of them are in category c. Then the prior probability of category c is

$$P(c)=\frac{A}{M}$$

And prior probability of other category is

$$p(c')=\frac{M-A}{M}$$

If a word w_n appears in category c for m_{wnc} and categories other than c for $m_{wnc'}$ times, then based on maximum likelihood estimation, the probability of this word in a specific category c is

$$p(w_n\,|\,c)=\frac{m_{w_nc}}{\sum_{n=1}^{N} m_{w_nc}}.$$

word in this tweet, any word w_{ik} condition on c or c' follows multinomial distribution. Therefore the probability of tweet d_i belongs to category c is

$$p(c\,|\,d_i)=\frac{p(d_i\,|\,c)\cdot p(c)}{p(d_i)}\propto\prod_{k=1}^{K}p(w_{ik}\,|\,c)\cdot p(c),$$

Posterior probability P(c| d_i) = P(c) *(likelihood probability of words from tweet)

If P(c| d_i) is larger than the probability threshold t, then di fit into category c, otherwise, di does fit into category c'. Other is only category if posterior probabilities of all categories are less than threshold value. Result of Naive Bayes (NB) classifier is stored in to naivebayes table.

The NB classifier algorithm is given in Algorithm 2. In this algorithm used collected tweets and categories as input for tweets classification. As a result, the output of algorithm NB classifier is classification of tweets into heavy study load (HSL), lack of social engagement (LOSE), negative emotions (NE), sleep problems (SP), soft-skill issues (SI) and other categories.

1) Algorithm NBclassifier (tweet, category)
2) Input: processed Tweets and categories
3) Output: Categories wise classification Tweets
4) // Total number of M tweets in the training set and A of them in category c.
5) // d_i is tweet, P(c) is prior probability, P (W_n/c) is likelihood probability.
6) // P (c| d_i) is posterior probability
7) begin

```
categories C: = $\{c_1, c2 \ldots. c_L\}$
for each category do
P(c) =A/M
end for
for each tweet do
divide $d_i$ into sub words $\{w_{1,} w_2, \ldots., w_N\}$
for each word do
P $(W_n/c)$ = count of word/ total count of words in category c
end for
P (c| $d_{i)}$ = P(c) * P($W_n$/c)
if P (c| $d_i$) > probability threshold then
$d_i$ does s fit into category c
end if
else
$d_i$ does fit into category c'
end else
end for
end
```

Algorithm 2: NB classifiers

III. RESULTS AND PERFORMANCE MEASURES

A. *Performance Measures*

To evaluate performance of system accuracy, precision, recall and f1-measure of four processes are used. There is a usually used Label based measure. Label based measures are calculated based on each category and then average over all category.

1) *Label Based Evaluation Measures*

Label based measures are calculated and average over each category. Create matrix in Table 1 for corresponding category c (Heavy study load).Similarly consider following matrix for each category.

	Tweet select by system	Tweet not select by system
Expected Tweet	True Positive(TRP) (actual heavy study load category tweets were correctly classified as heavy study load tweets)	False Negative(FAN) (heavy study load tweets s that were incorrectly marked as other category)
Not Expected Tweet	False Positive(FAP) (non- heavy study load tweets that were incorrectly classified as heavy study load)	True Negative(TRN) (all the remaining tweets are , correctly classified as non-Heavy study load tweets)

Table 1: Contingency table for Heavy study load

The sum of TRP, FAN, FAP and TRN is equal to total number of documents. Various metrics such as accuracy, precision, recall and F1 are used to measures the performance of classification algorithm. Then for one category c,

$$\text{Accuracy a}=\frac{\text{TRP+TRN}}{\text{TRP+TRN+FAP+FAN}}, \quad (1)$$

$$\text{Precision p}=\frac{\text{TRP}}{\text{TRP+FAP}}, \quad (2)$$

$$\text{Recall} \quad \text{r}=\frac{\text{TRP}}{\text{TRP+FAN}}, \quad (3)$$

$$\text{F1}=\frac{\text{2TRP}}{\text{2TRP+FAP+FAN}} \quad (4)$$

B. *Twitter data extraction result:*

Twitter tweets are collected using input query terms such as engineering Problem, DStudents problems, Alumini suggestions and lady Engineer. Using twitter API streamed tweets containing this query terms from January 2015 to July 2017.In total collected 16600 tweets. Table 2 shows summary of collected tweets.

Sr.no	Query term	count
1	engineeringProblem	12600
2	DStudentsproblems	2300
3	Aluminisuggestions	600
4	ladyEngineer	1100

Table 2: Tweets collection

1) *Categories Development Result.*

For categories development online Azure Machine Learning (AML) studio is used. From this studio choose LDA topic modeling algorithm. This algorithm applies on tweet dataset for to generate topics based on tweets content. Result of LDA algorithm for some tweets using 7 topic and 2-grams. It only generates topics. For topic name, collect topic wise tweets together using probability values greater than 0.4. For correct topic name and qualitative result perform content analysis on topic result generated by text analytics algorithm LDA. Tweets are categories in to heavy study load, lack of social engagement, negative emotions, sleep problems, soft-skill issues and other.

2) *Classification Result:*

From content analysis stage, a total of 600 #engineering Problems, #Alumini suggestions, and DStudents problems tweets an noted with 6 categories. Uses this tweets for training and testing. Observation shows that when the probability threshold value is 0.004 the performance is better than other threshold values. Table 3 shows the categories wise tweets count.

Category	Number of tweets
Heavy study load	114
Lack of social engagement	69
Negative emotion	69
Sleep problems	73
Soft skill issues	69
Other	344

Table 3: Number of tweets in each category for threshold value 0.004

Performance measures are calculated using equation 1, 2, 3 and 4. Table 4 shows category wise percentage for each measure. Found that accuracy for all categories is above 93% for threshold value 0.004 using Naïve Bayes classifier.

Category	Label a.	Label p.	Label r.	Label F1.
Heavy study load	93.24	68..42	95.12	79.59
Lack of social engagement	97.29	84.05	92.06	87.87
Negative emotion	97.29	86.95	86.95	88.23
Sleep problems	96.79	82.19	90.9	86.33
Soft skill issues	97.29	86.96	89.55	88.23
Other	96.45	75.00	75.16	75.00

Table 4: Category wise performance measure percentage

3) Detect Student problems from Offline dataset:

Naïve Bayes multi-label classifier is used to detect engineering student problems from offline dataset. There were 16000 tweets in offline dataset. Keep same threshold value for classifier for offline dataset. Table 5 shows number of tweets in each category for threshold value 0.004. So there is no extra human effort is needed when used classifier for classification.

Category	Number of tweets
Heavy study load	1309
Lack of social engagement	528
Negative emotion	316
Sleep problems	375
Soft skill issues	205

Table 5: Categories wise result of offline dataset for threshold value 0.004

IV. Conclusion

A proposed system works for twitter data extraction, tweets data preprocessing and tweets classification. Query terms engineering Problem, DStudents problems, Alumini suggestions and lady Engineer are very useful to collect relevant tweets. For category development topic modeling algorithm from ML studio is used. It gives categories wise tweets result. Content analysis is performed on ML studio result for category names and quality result. Naive Bayes multi label classifier algorithm performed on processed tweets for tweets classification. It gives the result in terms of accuracy above 93%, precision 85% and recall above 85%. In this workflow main problem of rural area engineering students such as soft skill issues is addressed. This proposed system result is very useful for educational policy makers to gain understanding of engineering students colleges' problems. It also provides a workflow for analyze Twitter data for educational purposes that overcomes the major limitations of traditional methods

References

[1] G.Siemens and P. Long, "Penetrating the fog; Analysis in learning and education,"Educause review, vol.48, no.5,pp.30-32, 2011.

[2] M. Clark, S. Sheppard, C. Atman, L. Fleming, R. Miller, R. Stevens, R. Streveler, and K. Smith, "Academic Pathways Study: Processes and Realities," Proc. Am. Soc. Eng. Education Ann. Conf. Exposition, 2008.

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

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

[5] 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.

[6] 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.

[7] Xin Chen, Mihaela Vorvoeanu, and Krishna Madhavan, "Mining social media data for understanding students learning experience," IEEE Transactions 2014.

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

[9] 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.