Analysis of Students Emotion for Twitter Data using Naïve Bayes and Non Linear Support Vector Machine Approachs

Ranjeeta Rana
Department of Computer Engineering,
D.Y.Patil College Of Engineering, Akurdi,
Savitribai Phule Pune University,
Pune, India
ranjeeta38@gmail.com

Mrs. Vaishali Kolhe
Department of Computer Engineering,
D.Y Patil College Of Engineering, Akurdi,
Savitribai Phule Pune University
Pune, India
vlkolhe@gmail.com

Abstract—Students' informal discussions on social media (e.g Twitter, Facebook) shed light into their educational understandings- opinions, feelings, and concerns about the knowledge process. Data from such surroundings can provide valuable knowledge about students learning. Examining such data, however can be challenging. The difficulty of students' experiences reflected from social media content requires human analysis. However, the growing scale of data demands spontaneous data analysis techniques. The posts of engineering students' on twitter is focused to understand issues and problems in their educational experiences. Analysis on samples taken from tweets related to engineering students' college life is conducted. The proposed work is to explore engineering students informal conversations on Twitter in order to understand issues and problems students encounter in their learning experiences. The encounter problems of engineering students from tweets such as heavy study load, lack of social engagement and sleep deprivation are considered as labels. To classify tweets reflecting students' problems multi-label classification algorithms is implemented. Non Linear Support Vector Machine, Naïve Bayes and Linear Support Vector Machine methods are used as multilabel classifiers which are implemented and compared in terms of accuracy. Non Linear SVM has shown more accuracy than Naïve Bayes classifier and linear Support Vector Machine classifier. The algorithms are used to train a detector of student problems from tweets.

Keywords-*social media, twitter, multi label classifiers, Naïve Bayes, Support Vector Machine*

*****

I. INTRODUCTION

Social networks have become very popular in recent years because of the increasing proliferation and affordability of internet enabled devices such as personal computers, mobile devices and other more recent hardware innovations such as internet tablets. In general, a social network is defined as a network of interactions or relations, where the nodes consist of actors, and the edges consist of the relations or interactions between these actors [1]. A generalization of the idea of social networks is that of information networks, in which the nodes could comprise either actors or entities, and the edges denote the relations between them. Social media sites provide great venues for students to share joy and struggle such as Twitter, Facebook, and Youtube, outlet emotion and stress and seek social support[28]. On various social media sites, students converse and share their everyday encounters in an informal and casual manner.

In machine learning, support vector machine is supervised learning model. It analyze the data and recognize the patterns for the classification and regression. SVM can efficiently perform the non linear classification using the kernel trick. Naïve Bayes classifier is the probabilistic model based on Bayes' Theorem. SVM is one of the most used and accurate classifiers in many machine learning tasks, and our comparison shows that Non Linear SVM exceeds Naïve Bayes and Linear SVM. The limitation of support vector approach lies in choice of kernel and it is high algorithmic complexity and extensive memory requirements of the required programming in the large-scale tasks[22]. SVM is a binary classifier, to do multi-class classification, pair-wise classifications is used(one class against all others, for all classes) [2]. In machine learning, the (Gaussian) radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification.

Students' ordinal ways provide vast amount of implicit knowledge and a whole new perspective for educational researchers and practitioners to understand students' experiences outside the controlled classroom environment. This understanding can inform institutional decision-making on interventions for unprotected students, enhancement of education quality, and thus increase student enrollment, maintenance, and success. The richness of social media data provides prospects to understand students' experiences but also raises methodological difficulties in making sense of social media data for educational purposes. Just imagine the sheer data volumes, the diversity of Internet slangs, the randomness of locations, and timing of students posting on the web as well as the complexity of students experiences [3]. Pure manual analysis cannot deal with the ever growing scale of data, while pure automatic algorithms usually cannot capture in depth meaning within the data [4].

The educational researchers usually have been using methods such as surveys, interviews, focus groups, class room activities to collect data related to students learning experiences. Therefore these methods are usually very time consuming, so cannot be duplicated or repeated with high frequency [5]. The balance of such studies is also usually limited. In addition, when prompted about their experiences, students need to reproduce on what they were thinking and doing sometime in the past, which may have become obscured over time.

The emerging field of learning analytics and educational data mining has focused on analyzing structured data obtained from course management system, classroom technology usage, or controlled online learning environments to inform

educational decision making [13]. However to the best of knowledge, there is no research found to directly mine and analyze student-posted content from uncontrolled spaces on the social web with clear goal of understanding students learning experiences. The goals are: 1)To demonstrate a workflow of social media data sense making for educational purposes, integrating both qualitative analysis and large-scale data mining techniques, 2)To discover engineering students informal discussions on Twitter, in order to understand subjects and difficulties students encounter in their learning experiences.

In educational environments there are many different types of data accessible for mining[29]. These data are exact to the educational area and so have intrinsic semantic information, relations with other data and multiple levels of meaningful order. The educational data and difficulties have some special characteristics that require the issue of mining to be treated in a different way. Although most of the traditional data mining techniques can be useful directly, others cannot and have to be adapted to the specific educational problem at hand[6,22].

II. LITERATURE REVIEW

Visualization of a social network can therefore be extremely useful to make people aware of their social context and to enable them to explore[35,36]. The main intent of information visualization is to represent an abstract information space in a dynamic way, so as to facilitate human interaction for exploration and understanding. The authors [19] analysed how more novel visualization techniques can be used to enhance various activities during the learning process: finding and understanding educational resources, collaboration with learners and teachers, (self-) reflecting about learners" progress, and designing learning experiences. The authors[19] illustrated analysis with example tools and visualizations [19].

Advanced educational technologies are developing rapidly and online MOOC courses are becoming more prevalent, creating an enthusiasm for the seemingly limitless data-driven possibilities to affect advances in learning and enhance the learning experience. For these possibilities to unfold, the expertise and collaboration of many specialists will be necessary to improve data collection, to foster the development of better predictive models, and to assure models are interpretable and actionable[8,23,25]. The big data collected from MOOCs needs to be bigger, not in its height (number of students) but in its width—more meta-data and information on learners' cognitive and self-regulatory states needs to be collected in addition to correctness and completion rates. This more detailed articulation will help open up the black box approach to machine learning models where prediction is the primary goal. Instead, a data-driven learner model approach uses fine grain data that is conceived and developed from cognitive principles to build explanatory models with practical implications to improve student learning[29].

Recent innovations in online education, including open online courses at various scales, in flipped classroom instruction, and in professional and corporate training have presented interesting questions about SLN[24,30]. Collecting, analyzing, and leveraging data about SLN lead to potential answers to these questions, with help from a convergence of modelling languages and design methods, such as social network theory, science of learning, and education information technology. This survey article overviews some of these topics, including prediction, recommendation, and personalization, in this emergent research area. The rapid deployment of Massive Open Online Courses (MOOCs) has created a surge in the global connectivity among students for educational purposes[13].

By definition, human-centered design relies on interaction with users. While interacting with users within industry can be challenging, fostering these interactions in a classroom setting can be even more difficult. This qualitative study explores the use of crowd-based design activities as a way to support student-user interactions online[7]. The authors motivated these online methods through a survey of 27 design instructors, identified common challenges of conducting student-user interactions in physical settings, including coordination constraints and geographical barriers to meeting in person. Authors then described their research through design approach to create and test 10 activities in a classroom setting, including using Twitter for need finding and using Reddit to brainstorm ideas with experts. Finally, authors presented an emergent framework outlining the design space for crowd-based design activities where students learn to use input from online crowds to inform their design work[9,40,41].

With the proliferation of smartphones and the increasing popularity of social media, people have developed habits of posting not only their thoughts and opinions, but also content concerning their where abouts. On such highly-interactive yet informal social media platforms, people make heavy use of informal language, including when it comes to locative expressions. Such usage inhibits the ability of traditional Natural Language Processing approaches to retrieve geospatial information from social media text. In this research, authors[25]: (1) develop a medium-scale corpus of "locative expressions" derived from a variety of social media sources; (2)benchmark the performance of a range of geoparsers over the corpus, with the finding that even the best-performing systems are substantially lacking; and (3) carry out extensive error analysis to suggest ways of improving the accuracy and robustness of geoparsers[25,26]. An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

The authors[2] represented the study of the life cycle of news articles are posted online. The authors described the interplay between website visitation patterns and social media reactions to news content. It show that use of hybrid observation method to characterize distinct classes of articles. It also find that social media reactions can help predict future visitation patterns early and accurately. It validate our methods using qualitative analysis as well as quantitative analysis on data from a large international news network, for a set of articles generating more than 3,000,000 visits and 200,000 social media reactions. It is possible to model accurately the overall traffic articles will ultimately receive by observing the first ten to twenty minutes of social media reactions[7]. Achieving the same prediction accuracy with visits alone would require to wait for three hours of data. It also describe significant improvements on the accuracy of the early prediction of shelf-life for news stories.

Communication between humans is extensively colored and strongly affected by emotions of the speakers. By nature, humans adjust their responses based on the actions of their dialogue partner in a certain emotional way – responding sadly if they're down, happily if they're nice, and angrily if they're rude[14]. This results in a dynamic and rich communication experience–an aspect yet to be completely replicated in human-machine dialogue. Human communication is naturally colored by emotion, triggered by the other speakers involved in the interaction. Therefore, to build a natural spoken dialogue

system, it is essential to consider emotional aspects, which should be done not only by identifying user emotion, but also by investigating the reason why the emotion occurred[15,36]. The ability to do so is especially important in situated dialogue, where the current situation plays a role in the interaction. In this paper[30], we propose a method of automatic recognition of emotion using support vector machine (SVM) of linear kernel and present further analysis regarding emotion triggers. Experiments were performed on an emotionally colorful dialogue corpus.

Supervised classification algorithms require annotated data to teach the machine, by example, how to perform a specific task[16]. There are generally two ways to collect annotations of a dataset: through a few expert annotators, or through crowdsourcing services (e.g., Amazon's Mechanical Turk). Many machine learning datasets are noisy with a substantial number of mislabelled instances. This noise yields sub-optimal classification performance. The authors study a large, low quality annotated dataset, created quickly and cheaply using Amazon Mechanical Turk to crowd source annotations[17]. The authors described computationally cheap feature weighting techniques and a novel non-linear distribution spreading algorithm that can be used to iteratively and interactively correcting mislabelled instances to significantly improve annotation quality at low cost. Eight different emotion extraction experiments on Twitter data demonstrate that our approach is just as effective as more computationally expensive techniques. The techniques used by authors saved considerable amount of time[38].

Sentiment analysis systems pursuit the goal of detecting emotions in a given text with machine learning approaches. These texts might include three kinds of emotions such as positive, negative and neutral. Entertainment oriented texts, especially movie reviews, contain huge amount of possible emotional information. In this study, authors aimed to represent each movie reviews by using small number of features. For this purpose, information gain, chi-square methods have been implemented to extract features for decreasing costs of calculations and increasing success rate[18]. In experiments, employed corpus includes Turkish movie reviews, support vector machine and naïve bayes had been employed for classification and F1 score was used for performance evaluation. According to the experimental results, support vector machine achieved 83.9% performance value while classification of movie reviews in two (positive and negative)categories and also authors obtained the 63.3% performance value while classification with support vector machine into three categories[37].

III. PROPOSED SYSTEM

The proposed system works on tweets related to engineering problems. In the system five labels are defined: heavy study load, lack of social engagement, negative emotions, sleep problem and diversity issues. The objective is to explore engineering students informal conversations on Twitter in order to understand issues and problems students encounter in their learning experiences. The tweets are loaded and processed by standard text mining procedure called Pre-processing[19].

The figure 1 shows the architecture of proposed system. The data is first collected then categories assigned to the tweets. The tweets is preprocess i.e stemming, stop word cleaning and tokenization. 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 preprocessing from text documents. However, on tokenization stream of text is break into words, phrases or symbols. The preprocessing is done tweets collected online. The tf-idf values are evaluated. The model is trained using multilabel classifier. The confusion matrix values are evaluated. The linear multi-label Support Vector Machine, non linear (RBF kernel) Support Vector Machine, Naive Bayes multi-label classifier are implemented and compared. The non linear SVM multi-label classifier shows more accuracy than linear multi-label SVM and Naïve Bayes [21,37]. The polarity of tweets are calculated i.e negative, very negative, neutral, positive and very positive by the three classifiers. Then uncategorized data is given to the trained detector as input which gives output as labelled data. The accuracy, true positive, false positive graphs of both multi-label classifier are shown. The procedure of the proposed system is as follow:

1) In the step one data collection is done from twitter.
2) The preprocessing and tfidf is calculated in the step 2.
3) Inductive Content analysis procedure is performed and categories are identified in step 3.
4) Naïve Bayes classifier, Linear SVM, Non Linear SVM are applied to dataset in order to demonstrate its application in detecting students problems is step 4.
5) The confusion matrix calculation is done and graphical analysis of classifiers is done to know the better accuracy among the classifiers is step 5, 6, and 7.

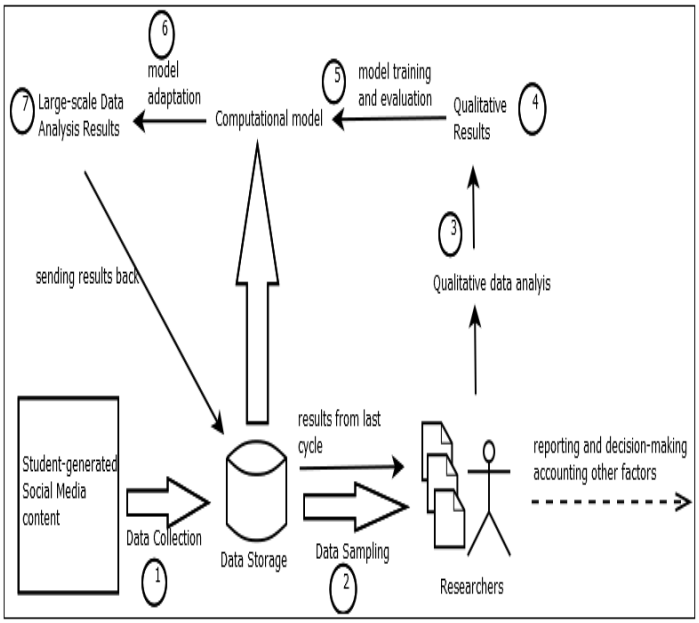

Figure 1. System Architecture

IV. METHODS

We have implemented three methods and compared results for better accuracy of tweets classification.

A. *Naïve Bayes Multi Label Classifier*

The basic concept is to assume independence among categories and train a binary classifier for each category. All kinds of binary classifier can be transformed to multilabel classifier using the one-versus-all heuristic. The following are the basic procedures of multi-label Naives Bayes classifier.

Suppose there are a total number of N words in the training document collection (each tweet is document) $W =$

$w_1, w_2, \ldots, w_N$ and total number of L categories $C = c_1, c_2, \ldots, c_L$.

If a word w_n appears in a category c for $m_{w_n c}$ times, and appear in categories other than c for $m_{w_n c}$ times , then based on the Maximum Likelihood Estimation, the probability of this in a specific category c is

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

Similarly, the probability of this word in categories other than c is

$$p(w_n|c') = \frac{m_{w_n c'}}{\sum_{n=1}^{N} m_{w_n c'}} \quad (2)$$

Suppose there are a total number of M documents in the training set and C of them are in category c. Then the probability of category c is

$$p(c) = \frac{C}{M'} \quad (3)$$

and the probability of other category c' is

$$p(c') = \frac{M-C}{M} \quad (4)$$

For a document d_i in the testing set, there are K words $W_{d_i} = \{w_{i1}, w_{i2}, \ldots, w_{iK}\}$ and W_{d_i} is a subset of W. The purpose is to classify this document into category c or not c. Assume independence among each word in this, document and any word w_{ik} conditioned on c and c' follows multinomial distribution. Therefore according to Bayes' Theorem, the probability that d_i belonged to category c is

$$p(c|d_i) = \frac{p(d_i|c).p(c)}{p(d_i)} \propto \prod_{k=1}^{K} p(w_{ik}|c).p(c) \quad (5)$$

and the probability that d_i belongs to categories other than c is

$$p(c'|d_i) = \frac{p(d_i|c').p(c')}{p(d_i)} \propto \prod_{k=1}^{K} p(w_{ik}|c').p(c') \quad (6)$$

$p(c|d_i) + p(c'|d_i) = 1$ normalize the latter two items which are propositional to $p(c|d_i)$ and $p(c'|d_i)$ to get real values of $p(c|d_i)$. If $p(c|d_i)$ is larger than the probability, then d_i belongs to category c, otherwise d_i belongs to c'. Then repeat this procedure for each category.

B. Support Vector Machine:

For Linear Support Vector Machine Given some training data D, a set of n points of the form

$$D = \{(x_i, y_i | x_i \in R^p, y_i \in \{1, -1\})\}_{i=1}^{n} \quad (7)$$

y_i is either 1 or -1, signifying the class to which the point x_i belongs. Each x_i is a p-dimensional real(R) vector. The aim is to find the maximum-margin hyperplane that divides the points having $y_i = 1$ from those having $y_i = -1$. The hyperplane can be written as the set of points x satisfying Maximum-margin hyperplane. Examples on the margin are called the support vectors.

$$w.x - b = 0 \quad (8)$$

where . denotes dot product and w the normal vector to the hyperplane. The parameter $\frac{b}{||w||}$ regulates the offset of the hyperplane from the origin along the normal vector w. If the training data are linearly separate, then two hyperplanes is selected in a way that they distinct the data and there are no points between them, and then attempt to exploit their distance. The area bounded by them is called the margin. The hyperplanes can be described by the equations

$$w.x - b = 1 \quad (9)$$

and

$$w.x - b = -1 \quad (10)$$

By using geometry, find the distance between these two hyperplanes is $\frac{2}{||w||}$, so minimize $||w||$. To prevent data points from falling into the margin, add the following constraint: for each i either

$$w.x_i - b \geq 1 \text{ for } x_i \text{ of first class} \quad (11)$$

or

$$w.x_i - b \leq -1 \text{ for } x_i \text{ of second class} \quad (12)$$

Using a Lagrangian, this optimization problem can be converted into a dual form can be expressed as

$$argmin_{w,b} max_{\alpha \geq 0} \{\frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i [y_i(w.x_i - b) - 1]\} \quad (13)$$

The Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient conditions for an optimal point. The Karush-Kuhn-Tucker condition implies that the solution can be expressed as a linear combination of the training vectors

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \quad (14)$$

Only a few α_i will be greater than zero. The corresponding x_i are exactly the support vectors, which lie on the margin and satisfy $y_i(w.x_i - b) = 1$.

For non linear support vector machine the resulting algorithm is formally similar, except that every dot product is replaced by a nonlinear kernel function. This allows the algorithm to fit the maximum-margin hyperplane in a transformed feature space. The transformation may be nonlinear and the transformed space high dimensional; thus though the classifier is a hyperplane in the high-dimensional feature space, it may be nonlinear in the original input space. The (Gaussian) radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification. The RBF kernel on two samples x and x' represented as feature vectors in some input space, is defined as

$$K(x, x') = \exp\left(-\frac{||x-x'||^2}{2\sigma^2}\right) \quad (15)$$

$||x - x'||^2$ may be recognized as the squared Euclidean distance between the two feature vectors. σ is a free parameter. An equivalent, but simpler, definition involves a parameter $\gamma = -\frac{1}{2\sigma^2}$.

$$K(x, x') = \exp(\gamma||x - x'||^2) \quad (16)$$

Since the value of the RBF kernel decreases with distance and ranges between zero (in the limit) and one (when x = x'), it has a ready interpretation as a similarity measure. The feature space of the kernel has an in finite number of dimensions; for $\sigma = 1$ its expansion is

$$\exp\left(-\frac{1}{2}||x - x'||^2\right) = \sum_{j=0}^{\infty} \frac{(x^T x')^j}{j!} \exp\left(-\frac{1}{2}||x||^2\right) \exp\left(-\frac{1}{2}||x'||^2\right) \quad (17)$$

A radial basis function (RBF) is a real-valued function whose value depends only on the distance from the origin, so that $\emptyset(x) = \emptyset(||x||)$ or alternatively on the distance from some other point c, called a center, so that $\emptyset(x, c) = \emptyset(||x - c||)$. Any function $\emptyset$ that satisfies the property $\emptyset(x) = \emptyset(||x||)$ is a radial function. The norm is usually Euclidean distance, although other distance functions are also

 ISSN: 2321-8169
3211 – 3217

possible. Radial basis functions are typically used to build up function approximations of the form

$$y(x) = \sum_{i=1}^{N} w_i \emptyset(||x - x_i||) \quad (18)$$

where the approximating function $y(x)$ is represented as a sum of N radial basis functions, each associated with a different center x_i and weighted by an appropriate coefficient w_i. The weights w_i can be estimated using the matrix methods of linear least squares, because the approximating function is linear in the weights.

V. RESULT

Twitter post is accessed by mentioning a access and token keys of site and query mentioned according to load the posts. The tweets are collected from twitter social media by twitter API. The Table 1 shows the values of TPR(True Positive Rate), SPC(Specificity or True Negative Rate), Precision, NPV(Negative Prediction Value), FPR(False Positive Rate), FDR(False Discovery Rate), FNR(False Negative Rate), Accuracy, F1 score(Harmonic mean of precision) for both multilabel classifiers. Then we compared linear multi-label support vector machine and Naïve Bayes multi-label classifier results. The table 2 defines the positive, very positive, very negative, negative and neutral values are calculated of tweets on the aspects lack of social engagement, heavy study load, negative emotions, sleep problem, diversity issue by three classifiers.

TABLE I. CONFUSION MATRIX VALUES OF NB, LINEAR SVM AND NON LINEAR SVM ON TWEETS

Confusion Matrix	NB(Naïve Bayes)	Linear Support Vector Machine	Non Linear Support Vector Machine
SPC	92%	97.2%	98.1%
Precision	68.5%	88.8%	92.5%
TPR	68.5%	88%	92.5%
NPV	92.1%	97.2%	98.1%
FPR	7.8%	2.7%	1.8%
FDR	31.4%	11.11%	7.4%
FNR	31.4%	11.11%	7.4%
Accuracy	87.4%	84.4%	97.3%
F1 Score	91%	88.88%	92.5%

TABLE II. POLARITY VALUES OF LABELS ON TWEETS USING NON LINEAR SVM

Aspect	Very Negative	Negative	Neutral	Positive	Very Positive
Lack of Social Engagement	0.0	0.705	0.156	0.137	0.0
Heavy Study Load	0.0	0.787	0.148	0.063	0.0
Negative Emotion	0.0	0.811	0.150	0.037	0.0
Sleep Problem	0.196	0.803	0.137	0.392	0.0
Diversity Issue	0.0	0.820	0.134	0.044	0.0

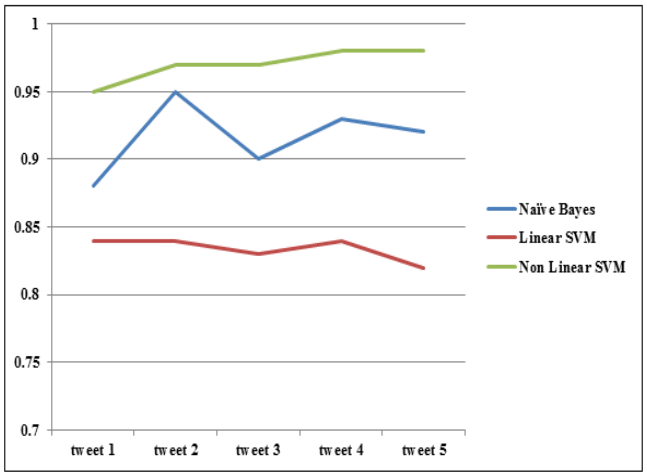

Figure 2. Accuracy Graph Analysis of Naïve Bayes, Linear Support Vector Machine and Non Linear Support Vector Machine

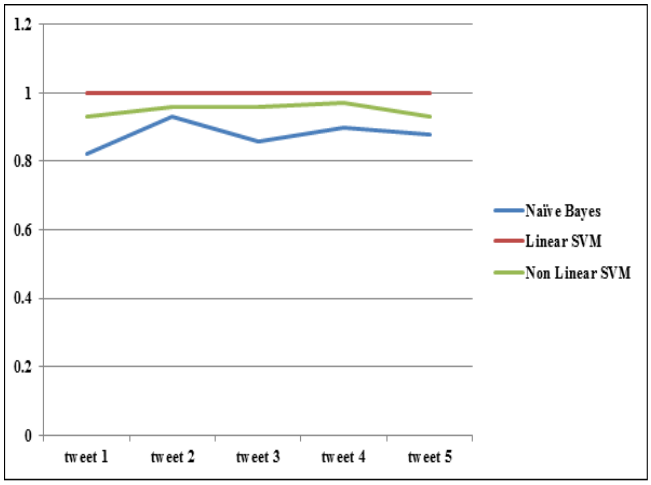

Figure 3. F1 Score Graph Analysis of Naïve Bayes, Linear Support Vector Machine and Non Linear Support Vector Machine

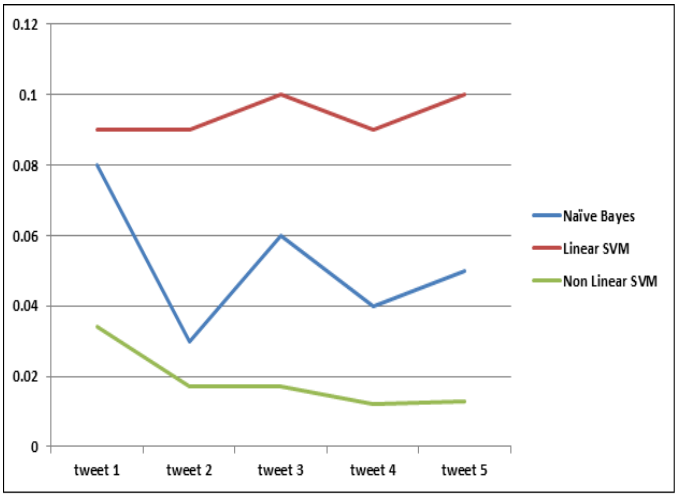

Figure 4. False Positive Graph of Naïve Bayes, Linear Support Vector Machine and Non Linear Support Vector Machine

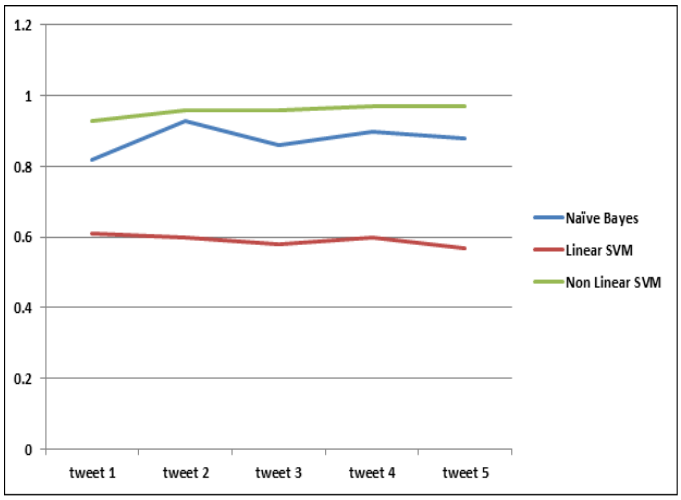

Figure 5. True Positive Graph of Naïve Bayes, Linear Support Vector Machine and Non Linear support Vector Machine

VI. Conclusion And Future Work

It provides the work flow for analyzing social media data for educational purposes that overcomes the major limitations of both manual qualitative analysis and large scale computational analysis of user-generated textual content. The results are compared of Support Vector Machine and Naïve Bayes algorithms and evaluated a multi-label classifier to detect engineering student problems. This study explores space on twitter in order to understand engineering students experiences, integrating both qualitative methods and large-scale data mining techniques. The comparison result has shown more accuracy of Non Linear SVM than Naive Bayes classifier and Linear SVM.

Future work could analyze students' generated content others than texts(e.g. images and videos), on social media sites other than Twitter(e.g. Facebook, Tumbler and YouTube). Future work can be done to design more sophisticated algorithms in order to reveal the hidden information in the "long tail". The manipulation of personal image online may need to be taken into considerations in future work.

Acknowledgment

The authors would like to thank the researchers as well as publishers for making their resources available and the teachers for their guidance. We also thank the organizations for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to all friends and family members.

References

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

[2] Carlos Castillo, Mohammed El-Haddad, Jurgen Pfeffer, and Matt Stempeck, 'Characterizing the life cycle of online news stories using social media reactions', In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, pages 211-223, ACM, 2014.

[3] Yue Gao, Fanglin Wang, Huanbo Luan, and Tat-Seng Chua, 'Brand data gathering from live social media streams', In Proceedings of International Conference on Multimedia Retrieval, page 169, ACM, 2014.

[4] Abdullah Gok, Alec Waterworth, Philip Shapira, 'Use of Web Mining in Studying Innovation', Scientometrics, pp. 653-671, Springer, 2015.

[5] Julie S Hui, Elizabeth M Gerber and Steven P Dow, 'Crowd-based Design Activities: Helping Students connect with Users Online', In Proceeding of the 2014 conference on Designing Interactive Systems, pp. 875-884, ACM, 2014.

[6] Joris Klerkx, Katrien Verbert and Erik Duval, 'Enhancing learning with visualization techniques', In Handbook of Research on Educational Communications and Technology, pp 791-807, Springer, 2014.

[7] Fei Liu, Maria Vasardani and Timothy Baldwin, 'Automatic Identification of Locative Expressions from Social Media Text: A comparative analysis', In Proceedings of the 4th International Workshop on Location and the Web, pp 9-16, ACM, 2014.

[8] Nguyen, Thin and Phung, Dinh and Adams, Brett and Venkatesh, Svetha, 'Mood Sensing from Social Media Texts and its Applications', Knowledge and information systems, vol. 39, number 3, pp. 667-702, Springer, 2014.

[9] Opuszko, Marek and Ruhland, Johannes, 'Classification Analysis in Complex Online Social Networks Using Semantic Web Technologies', Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONA 2012), pp. 1032-1039, IEEE Computer Society, 2012.

[10] Popescu, Elvira, 'Providing collaborative learning support with social media in an integrated environment', World Wide Web, vol. 17, no.2, pp. 199-212, Springer, 2014.

[11] Rabbany, Reihaneh and Elatia, Samira and Takaffoli, Mansoureh and Zaiane, Osmar R, 'Collaborative Learning of Students in Online Discussion Forums: A Social Network Analysis Perspective', In Educational Data Mining, pp. 441-466, Springer 2014.

[12] Ribarsky, William and Xiaoyu Wang, Derek and Dou, Wenwen, 'Social media analytics for competitive advantage', Computers & Graphics, vol. 38, pp. 328-331, Elsevier, 2014.

[13] Romero, Cristobal and Ventura, Sebastian, '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, IEEE, 2010.

[14] Shridhar, Saajan and Gupta, Ankur and Shridhar, Swapan, 'Improving Student Engagement in Higher Education: An Experiment with a Facebook Application in India', International Journal of Computer Science:Theory, Technology and Applications (IJCS), vol. 3, no. 1, 2014.

[15] Tang, Jie and Zhang, Yuan and Sun, Jimeng and Rao, Jinhai and Yu, Wenjing and Chen, Yiran and Fong, Alvis Cheuk M, 'Quantitative study of individual emotional states in social networks', Affective Computing, IEEE Transactions on, vol. 3, no. 2, pp. 132-144, IEEE, 2012.

[16] Thovex, Christophe and Trichet, Francky, 'Opinion mining and semantic analysis of touristic social', Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on, pp. 1155-1160, IEEE, 2013.

[17] Tuarob, Suppawong and Tucker, Conrad S and Salathe, Marcel and Ram, Nilam, 'An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages', Journal of biomedical informatics, Elsevier, 2014.

[18] Vongsingthong, Suwimon and Wisitpongphan, Nawaporn, 'Classification of university students' behaviors in sharing information on Facebook', Computer Science and Software Engineering (JCSSE), 2014 11th International Joint Conference on, pp. 134-139, IEEE, 2014.

[19] Wu, Xindong and Zhu, Xingquan and Wu, Gong-Qing and Ding, Wei, 'Data Mining with Big Data',Knowledge and Data Engineering, IEEE Transactions on, vol. 26, no. 1, pp. 97-107, IEEE, 2014.

[20] Chen, Xin and Madhavan, Krishna and Vorvoreanu, Mihaela, 'A Web-Based Tool for Collaborative Social Media Data Analysis',Cloud and Green Computing (CGC), 2013 Third International Conference on, pp. 383-388, IEEE, 2013.

[21] Yakushev, Andrei and Mityagin, Sergey, 'Social Networks Mining for Analysis and Modeling Drugs Usage," Procedia Computer Science, vol. 29, pp. 2462-2471, Elsevier, 2014.

[22] Saleh Alsaleem, "Automated Arabic Text Categorization Using SVM and NB", International Arab Jornal of e-Technology, Vol.2, No. 2, June 2011.

[23] A. Sopharak, B. Uyyanonvara, S. Barman, "Comparing SVM and Naïve Bayees Classifier for Automatic Microaneurysm Detections", World Academy of Science, Engineering and Technology, International Journal of Computer, Control, Quantum and Information Engineering vol. 8, No. 5, 2014.

[24] Entezari-Maleki, Reza and Rezaei, Arash and Minaei-Bidgoli, Behrouz, "Comparison of classification methods based on the type of attributes and sample size", Journal of Convergence Information Technology, vol. 4, no. 3, pp.94-102, 2009.

[25] Bekkali, Mohammed and Lachkar, Abdelmonaime, "Arabic Tweets Categorization Based on Rough Set Theory", International Journal of Computer Science and Information Technology, vol. 6, no.6, 2014.

[26] Zhu, Linhong and Galstyan, Aram and Cheng, James and Lerman, Kristina, "Tripartite graph clustering for dynamic sentiment analysis on social media", Proceedings of the 2014 ACM SIGMOD international conference on Management of data, pp. 1531-1542, ACM, 2014.

[27] Zhang, Pu and He, Zhongshi and Tao, Lina, "A Study of Dependency Features for Chinese Sentiment Classification", Journal of Software, vol. 9, no. 11, pp. 2877-2885, 2014.

[28] Altrabsheh, Nabeela and Cocea, Mihaela and Fallahkhair, Sanaz, "Learning sentiment from students' feedback for real-time interventions in classrooms", Adaptive and Intelligent Systems, pp. 40-49, Springer, 2014.

[29] Opuszko, Marek and Ruhland, Johannes, 'Classification Analysis in Complex Online Social Networks Using Semantic Web Technologies', Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONA 2012), pp. 1032-1039, IEEE Computer Society, 2012.

[30] Hasan, Maryam and Rundensteiner, Elke and Agu, Emmanuel, "EMOTEX: Detecting Emotions in Twitter Messages", Academy of Science and Engineering (ASE), USA, ASE 2014.

[31] El-Makky, Nagwa and Nagi, Khaled and El-Ebshihy, Alaa and Apady, Esraa and Hafez, Omneya and Mostafa, Samar and Ibrahim, Shimaa, "Sentiment Analysis of Colloquial Arabic Tweets", Academy of Science and Engineering, USA, 2015.

[32] Anjaria, Malhar and Guddeti, Ram Mohana Reddy, "Department of Information Technology, National Institute of Technology Karnataka, Surathkal, Mangalore-575025, India", Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference , pp. 1-8, IEEE, 2014.

[33] Ram, Sudha and Zhang, Wenli and Williams, Max and Pengetnze, Yolande, "Predicting Asthma-Related Emergency Department Visits Using Big Data", IEEE, 2015.

[34] Altrabsheh, Nabeela and Cocea, Mihaela and Fallahkhair, Sanaz, "Predicting learning-related emotions from students' textual classroom feedback via Twitter", International Educational Data Mining Society, 2015.

[35] Mohammad, Saif M and Kiritchenko, Svetlana, "Using Hashtags to Capture Fine Emotion Categories from Tweets", Wiley Online Library, Computational Intelligence, 2014.

[36] Lubis, Nurul and Sakti, Sakriani and Neubig, Graham and Toda, Tomoki and Purwarianti, Ayu and Nakamura, Satoshi, "Emotion and its triggers in human spoken dialogue: Recognition and analysis", Proc IWSDS, 2014.

[37] Akba, F{\i}rat and U{\c{c}}an, Alaettin and Sezer, Ebru Ak{\c{c}}ap{\i}nar and Sever, Hayri and Alsaffar, Ali and Deogun, Jitender and Raghavan, Vijay and Sever, Hayri and Alsaffar, Ali and Deogun, Jitender and others, "Assessment of feature selection metrics for sentiment analyses: Turkish movie reviews", vol. 191, pp. 302-309, Sringer, 2014.

[38] Martineau, Justin and Chen, Lu and Cheng, Doreen and Sheth, Amit P, "Active Learning with Efficient Feature Weighting Methods for Improving Data Quality and Classification Accuracy", Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics,2014.

[39] Li, Jenny S and Monaco, John V and Chen, Li-Chiou and Tappert, Charles C, "Authorship Authentication Using Short Messages from Social Networking Sites", e-Business Engineering (ICEBE), 2014 IEEE 11th International Conference on, pp. 314-319, IEEE, 2014.

[40] Hagen, Matthias and Potthast, Martin and B{\"u}chner, Michel and Stein, Benno, "Twitter sentiment detection via ensemble classification using averaged confidence scores", Advances in Information Retrieval, pp. 741-754, Springer, 2015.

[41] Patra, Dr PSK and other, "Classification of Questions in Micro-blogging Environment Using Support Vector Machine", vol. 3, no. 4, pp. 540-544, IJRCCT, 2014.