International Journal of System and Software Engineering
7 (1), June 2019, 13-19
http://www.publishingindia.com/ijsse

Dynamic Induction Model for Student's Behavior Analysis

Sharmishta Desai
School of Computer Engineering and Technology, MITWPU, Pune, Maharashtra, India.
Email: sharmishta.desai@mitwpu.edu.in

Abstract: The volume of data is growing rapidly due to the usage of social sites like twitter, facebook etc. 80% of the college students spend their maximum time on social media. They share their views, feelings, emotions on it. This massive data is useful for institutes for getting feedback about any student or services provided by them. This feedback will help institutes to provide proper mentoring to students or to take any corrective action which will improve quality of service. The use of machine learning algorithms for analyzing this data will add more knowledge into the knowledge of institutes. Decision tree algorithm provides visual representation of data which is useful for social media data analysis. Traditional machine learning algorithms like C4.5 or CART have a limitation of memory size because they store all data on memory for building a model. So, these algorithms are not suitable for large volume of data. These algorithms performs best if the size of data is small but if size of data increases the same algorithms shows poor results. In this paper, we have used Hoeffding tree for large volume of data and proved with results that Hoeffding tree performs best against other Machine learning algorithms. Other algorithms like SVM, Naïve Bayes, Decision Tree C4.5 work well if the data set is small but their performance degrades if data size increases. To increase accuracy, we have used different classifiers at leaf level and analyzed different split criteria's. We have collected dataset from twitter social site. Different phases of social media data mining are also explained in detail.

Keywords: Decision trees, Hoeffding trees, Social media data.

I. Introduction

In a survey conducted by the Association for University and College Counselling Centre Directors, 36.4% of college students reported that they experienced some level of depression [32]. This depression is because of study load, fear of dropping out from college, exam pressure, peer pressure or ragging etc. College students are very much active social sites like facebook, twitter, you tube etc. As per the survey conducted by Pew Research Center, 72% of high school and 78% of college students spend time on Facebook, Twitter, Instagram, etc. These numbers indicate how much the student community is involved in this virtual world of social networking [35]. They share their feelings of joy, anger or frustration on these sites. This information can be used by colleges to understand their student's behaviour for mentoring purposes. This large amount of data is helpful for institutes to improve their teaching as well as mentoring quality. As this data is large in volume and diverse in formats, we need a machine learning approach which will take care of volume, velocity and variety. Naive Bayes, C4.5, CART, SVM are traditional machine learning algorithms which performs best if the size of data is small. We have implemented a dynamic induction model based on Hoeffding tree and used it to analyse engineering students behaviour. We have created one hashtag, '#Engineering Student's Problems' and shared among various Engineering students. The remainder of the paper is organized as follows:

Next section provides literature review. In Section III, proposed methodology is explained. Section IV gives experimental results.

II. Literature Review

In literature many authors have used different machine learning algorithms to analyse time changing big data. In [1], author has used neural network to analyze behaviours of customers using social media data set. In [2], author has explained a way to find link between to users social media like twitter or facebook using machine learning algorithm. In [3, 4, 5], authors have explained big data architecture, challenges etc. In [6, 7, 8, 9, 10, and 11], different approaches are explained to analyse social media data. Usage of SVM for data classification is explained in [12, 13, and 19]. Efficient usage of decision tree for social media data mining is given in [14, 17, and 18]. In [15], author has used SVM with differential evolution algorithm for classifying data. In work done by Isvani Frıas-Blanco, Jose del Campo-Avila [29], moving average method is suggested which is used for Online and Non-Parametric Drift Detection Methods Based on Hoeffding Bounds. Short-Term Load Forecasting Based on Big Data Technologies by Pei Zhang [31], explain decision tree framework to forecast short term load like electricity. Petra Perner has explained how decision tree induction is suitable than

traditional methods in his paper titled "Decision Tree Induction Methods and their Applications to Big Data". Also there is lot of work available on social media data analysis. Characteristics of social activities and patterns of communication in Twitter are studied by Naaman *et al.* [20]. Davidov *et al.* [21] have used hash tags and other sentiment labels for sentiment analysis. An effective and efficient followee recommender system built by Hannon *et al.* [22]. Methods to recommend influential users proposed by Kwak *et al.* [23]. In [26], ranking based recommendation system is explained. Twitter use within and across organizations and geographic markets comparison is proposed by Burton *et al.* [24]. Map reduce implementation of C4.5 is explained in [28]. Kim *et al.* [25], explained how to maximize the outcomes of SMM through Word-of-Mouth (WOM) marketing by identifying the core group of users. In [16], author has explained how to extract knowledge using decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing. Distributed implementation of support vector machines is proposed by [10]. In [30], combination of Naïve Bayes with decision tree is explained. In [27] author has used multi label classifier to locate the relevant data and Twitter hashtags. They collected 25,284 tweets using the hashtag # engineering Problems over a period of 14 months, and a second data set of 39,095 tweets using the geo-code (longitude and latitude) of Purdue University, West Lafayette. In [33], authors proposed partial supervised learning for HDP which enables HDP to make use of partial known knowledge to guide the model learning process. This partial learning enables HDP which is aimed at solving clustering problems to tackle classification problems and meanwhile partial supervised learning helps improve the classification accuracy. They applied the proposed partial supervised learning for HDP to classify posts (micro-blogs) in an educational environment. In paper [34] authors proposes a novel application of text categorization to identify relevant and irrelevant micro-blogging questions asked in a classroom. Several modelling approaches and several weighting or pre-processing configurations are studied for this application through extensive experiments.

III. Proposed Methodology

We have created hashtag '#engineeringstudntproblems' and shared among MIT College students. We have collected more than 5000 tweets. Also we have collected tweets of different longitude and latitude through NodeXL graph with hashtag "#studentsproblems'. The overall architecture is given in Fig. 1.

As the data is dynamic, we need an approach which is based on local optima model. This model does not require whole data set to be present on a memory as well as it can be modified as and when required. To avoid missing, biased and concept drift data, we have concentrated on preprocessing technique. Proposed technique is based on distributed architecture which contains all steps from data preprocessing to tree optimization. This technique is based on horizontal splitting of samples.

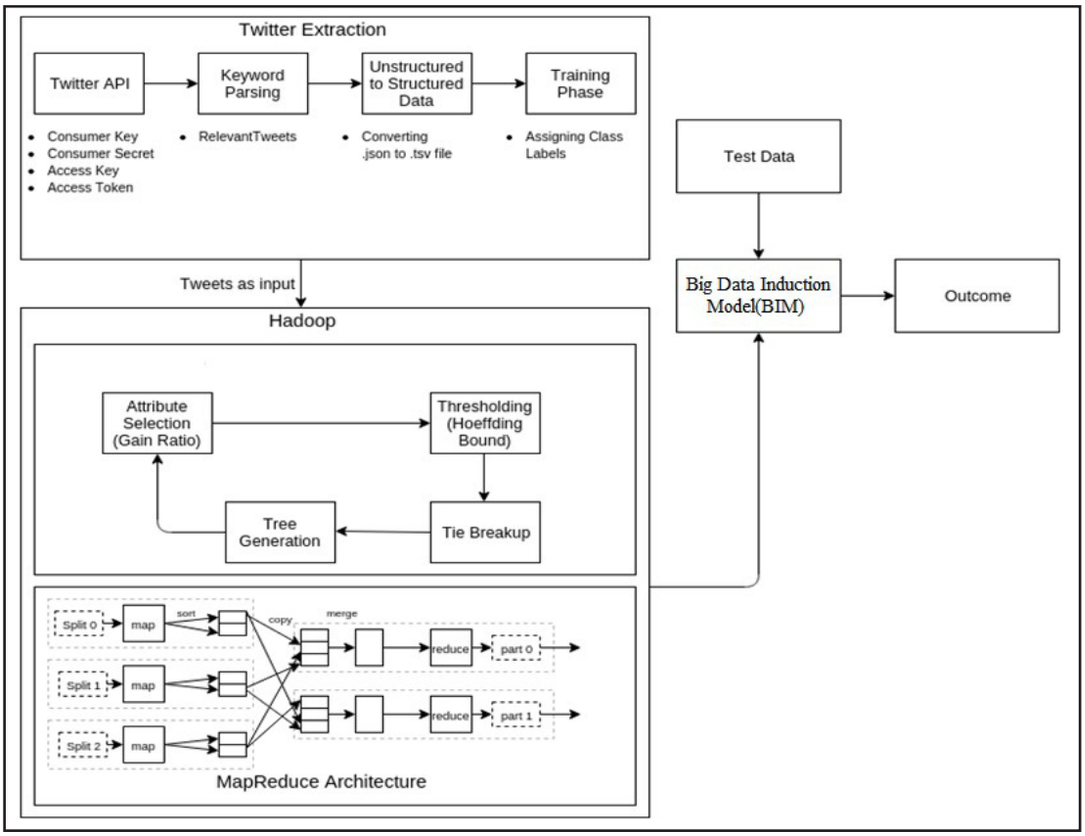

Fig. 1: Proposed Architecture for Tweets Extraction and Classification

A. Twitter Extraction and Labeling

Tweets are extracted using twitter API. Tweets are labeled to different classes based on keyword parsing and weightages assigned to different adjectives or smiles. Different classes used are positive(C+), negative(C-) and neutral(Cn). Positive class label indicates mentoring is not required. Negative class label indicates mentoring is required while neutral indicates can't say anything.

Sometimes tweets do not convey meaningful information such tweets are removed from dataset to reduce training time. Unstructured tweets data set is converted into structured data set as given below. Example of Student related tweet shown in Table I.

TABLE I: TWEET SENTIMENT CATEGORIES

Sentiment	Tweet
Positive	Hurray got distinction Now its time to enjoy!!!Got break
Negative	Anyone else frustrated with project work? Oh God, save me from hell submissions
Neutral	Submissions are good but not gives no output

To do proper data pre-processing and cleaning, following algorithm is proposed.

Input:- X1,X2,…… (Non Ending Stream of tweets)

Output:-X1,X2,…… (Processed Stream of tweets)

Algorithm:

For all words xi ∈ {x1, x2,……}

For all Xi ∈ {' ' }

Compute frequent fi ∈ {f1, f2,,……}

Assign weights Wi ∈ {W1, W2,,……} based on adjectives & phrases.

Calculate total weight

$$w = \sum_{i=1}^{k} fiwi$$

If w>0 then label(C+)

If w<0 then label(C-)

If w=0 then label (Cn)

B. Feature (Attribute) Selection

After labeling tweets, different attributes like data, time, longitude, latitude are extracted. This is helpful to identify which region students face problems. Also we can find in which duration students are more frustrated. Proper selection and construction of features is a critical task. It affects the result of machine learning algorithm execution.

Features are evaluated based on their information gain. The feature or attribute having larger information gain considered as a node in decision tree. If several features are representing same information then some other features are combined or some are deleted from this. Every feature is ranked according to its information gain. Features are selected in a subset. The number of attributes in one subset is defined by thresholding.

$$H(x) = \sum_{i=1}^{n} p(Xi) \log p(Xi)$$

Info_Gain(X) = H(X1) - H(X2) Where H(X1) - Entropy before split and H(X2)-Entropy after split.

The thresholding method is explained below.

C. Thresholding

Thresholding is required to limit features into subset. Thresholding will set the number of features in a data set which are sufficient for finding information gain on that set. For finding threshold there is no thumb rule. Based on trial and error in learning phase threshold value can be calculated.

Algorithm:

Compute Entropy for all Ai,

$$E(Ai) = \sum_{i=1}^{t} P(Ai) \log(P(Ai)$$

Where t are the number of samples for Attribute Ai.

Compute Information Gain for all Ai,

$$I\,(Ai) = E1\,(Ai) - E2\,(Ai)$$

Where E1 (Ai) is entropy before split and E2 (Ai) is entropy after split.

Select attribute Ai, I (Ai) > = max (At) where t is from 1 to n.

Select Ai as splitting attribute.

Ai→left = all Ati; I (Ai) > I (Ati)

Ai→ right = all Ati; I (Ai) <I (Ata)

D. Tie Breakup

There are some attributes for which there is no difference in information gain in two attributes. Such attributes are called as tie attributes. When such attributes occur then one out of two is selected.

Traverse all Ai ε GT

Compute Information Gain, I

For attributes A1, A2

$$\text{If } I\,(A1) = = I\,(A2)$$

Calculate HB using following formula,

$$HB = \sqrt{\frac{R2\text{In}\left(\frac{1}{\delta}\right)}{2n}}$$

If I (A1)-I (A2) > HB
Omit A2 from the tree
Readjust the tree

E. Creating Social Tree

Convert the social media data into social graphs.

Pre-process raw data and social graphs so that they become suitable for applying ML algorithms.

Read social media data into leaf node.

Calculate information gain on attributes using following formula:

$$H(X) = \sum_{i=1}^{n} p(Xi) \log p(Xi)$$

Info_Gain(X) = H(X1) - H(X2) Where H(X1) - Entropy before split and H(X2)-Entropy after split.

Select the attribute with highest gain using Hoeffding bound (HB) criteria.

H (Xa) - H (Xb) > HB then select attribute Xa or select Xb.

HB is calculated using following formula:

$$HB = \sqrt{\frac{R2\ln(\frac{1}{\delta})}{2n}}$$

Split the node and add two leaf nodes.

Xa→ left = for all attributes Xi, H (Xi) < H (Xa)

Repeat above steps till $X \neq \Theta$

IV. Results

Experimental results are generated using Weka 3.7.11 Java library and Eclipse Europa on Windows platform. Engineering students tweets extracted from twitter is used for experimentation. For student's data set decision tree, Naïves Bayes and SVM algorithms are evaluated.

Following steps are followed to extract twitter data.

1. Create a script having a unique id or twitter account to extract data.
2. Use twitter API to extract the tweets, the extracted tweets are the tweets of the current day. We can also narrow our search down to extract tweets related to a particular domain.
3. Once the data has been extracted, it is organized based on required functions in a database.
4. Social parameters tell us about the closeness between two users, i.e., user affinity and the influence of one user over other, i.e., user Influence. Social parameters are instantiated by creating a graph of social network and observing the interactions between nodes.
5. Location parameters are based on the current location of the user which is narrowed down by time zones and the places where the user checks-in.
6. Time parameters are just the regency of a mention.
7. Machine learning algorithm is executed on above created dataset.

The result of different machine learning algorithms on student's data is shown below.

Table II: Comparison of Different Machine Learning Algorithms Performance

Algorithm Name	Correctly Classified	Incorrectly Classified	F-measure	ROC	Time in Seconds
Decision Tree C4.5	2948	1679	0.496	0.499	0.03
Decision Tree CART	2948	1679	0.496	0.499	0.49
Random Forest	2948	1679	0.496	0.5	0.45
Naive Bayes	2948	1679	0.499	0.496	0.02
SVM	2948	1679	0.496	0.5	0.19
Hoeffding Tree	2962	1665	0.596	0.582	0.09

The results given in above Table II are represented in a graph given below.

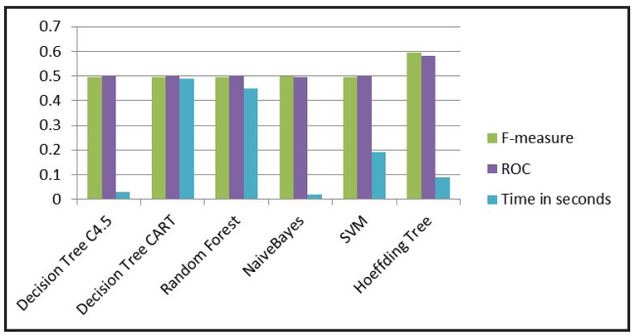

Fig. 2: Comparison of Performance of ML Algorithms

In above graph we can observe Hoeffding tree classify data more accurately. In above Fig. 2, we can see ROC and F-measure for Hoeffding Tree is greater than other algorithms. More is the ROC and F-measure, more accurate is the algorithm. Time required for Hoeffding Tree is less in comparison with other decision tree algorithm but it is more in comparison with Naive Bayes and SVM (Support Vector Machines).

Next all decision tree algorithms are applied on large data set. It is shown in Table III. These results show that C.4.5 works well till 50000 attributes. Hoeffding tree works well though number of attributes are beyond 50000. Though data increases; accuracy of hoeffding tree remains linear that is 74% while other algorithms fails due to their memory requirement.

TABLE III: COMPARISON OF DIFFERENT DECISION TREE ALGORITHMS

Algorithms	No. of Attributes	Time (in Seconds)	Accuracy (%)
C4.5	10000	0.28	72.74
Random Forest	10000	0.47	70.59
Hoeffding Tree	10000	0.17	73.83
C4.5	50000	1.45	73
Random Forest	50000	Fails	Fails
Hoeffding Tree	50000	0.78	73.662
C4.5	100000	Fails	Fails
Random Forest	100000	Fails	Fails
Hoeffding Tree	100000	1.56	73.9
C4.5	1000000(1GB)	Fails	Fails
Random Forest	1000000(1GB)	Fails	Fails
Hoeffding Tree	1000000(1GB)	74.78	74

Accuracy of hoeffding tree is more improved by using different classifiers at leaf level. It is used to remove uneven classification at leaf. We have used Naïve Bayes, Adaptive Naïve Bayes and Majority class classifiers as leaf level.

It is shown in Table IV.

TABLE IV: COMPARISON OF CLASSIFIERS USED AT LEAF LEVEL

Algorithms Used at Leaf Level	No. of Attributes	Correctly Classified	Incorrectly Classified	Time in seconds	Accuracy (%)
Majority Class	10000	1050	8950	0.13	10.5
Naïve Bayes	10000	7383	2617	0.05	73.83
Naïve Bayes Adaptive	10000	7383	2617	0.14	73.83

Naïve Bayes classifier at leaf level gives good accuracy in less time as compared to other classifiers as shown in Fig. 3 & Fig. 4.

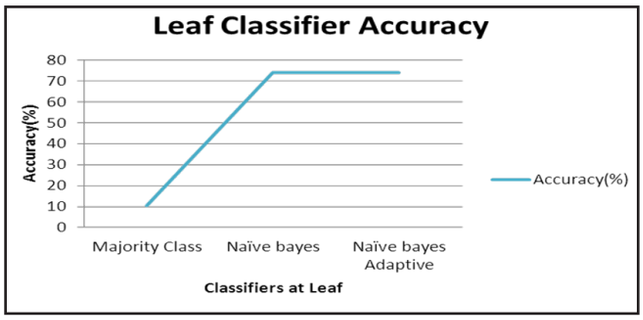

Fig. 3: Comparison of Accuracy of Leaf Classifier

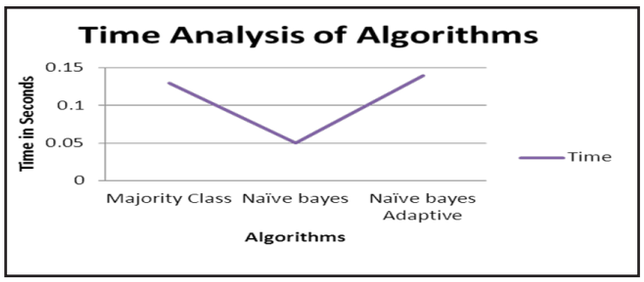

Fig. 4: Comparison of Learning Time of Leaf Classifier

Instead of entropy, information gain and Gini index are used as split criteria at attribute. Information gain and Gini index select the attribute with fewer splits which helps to avoid over fitting. It is useful to increase accuracy of algorithm. The results are shown in Table V and Fig. 5.

TABLE V: COMPARISON OF SPLIT CRITERIA AND CLASSIFIERS USED AT LEAF LEVEL

Algorithms Used at Leaf Level	Spilt Criteria	No. of Attributes	Time in Seconds	Accuracy (%)
Naïve Bayes Adaptive	Gini Index	10000	0.05	85.71
	Info Gain	10000	0.06	86.68
Naïve Bayes	Gini Index	10000	0.03	85.75
	Info Gain	10000	0.02	87.04
Majority Class	Gini Index	10000	0.02	84
	Info Gain	10000	0.02	84

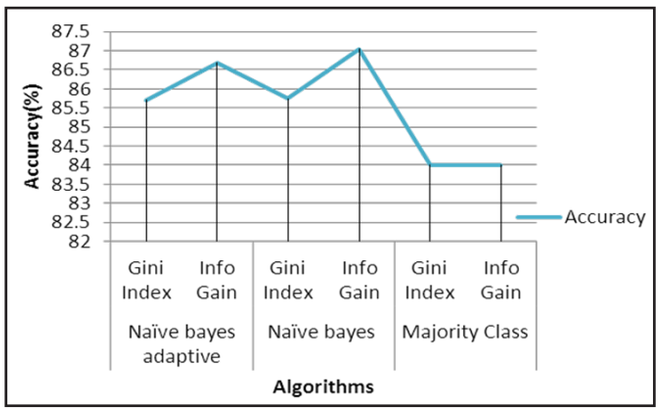

Fig. 5: Comparison of Split Criteria and Classifiers Used at Leaf Level

V. CONCLUSION

Due to popularity of social media, students tend to share their emotions, feelings on social media. This data is useful for institutes to find student problems and to mentor students in right direction. Machine learning algorithms are very much useful for doing this analysis. In this paper we have collected data from

twitter using NodeXL and preprocessed by executing python script. Then we have filtered it according to our requirement and then it is used with different Machine Learning (ML) algorithms. C4.5, CART, SVM are traditional machine learning algorithms which performs best if the size of data is small but if size of data increases the same algorithms shows poor results and after some limit they fails. It is found that Hoeffding tree algorithm's performance is best. Also this algorithm takes less execution time as compared to other decision tree algorithms.

Acknowledgements

We would like to thank MIT College of Engineering, Pune, India for providing infrastructure to do this research. Also we like to thank Savitribai Phule Pune University, Pune, India for sponsoring our research topics under BCUD research grant scheme.

References

[1] B. Zheng, K. Thompson, S. S. Lam, S. W. Yoon, and N. Gnanasambandam, "Customers' behavior prediction using artificial neural network," *Proceeding of the 2013 Industrial and Systems Engineering Research Conference*, pp. 700-709, 2013.

[2] O. J. Mengshoel, R. Desai, A. Chen, and B. Tran, "Will we connect again? Machine learning for link, prediction in mobile social networks," *Eleventh Workshop on Mining and Learning with Graphs*, Chicago, Illinois, USA, 2013.

[3] C. Yadav, S. Wang, and M. Kumar, "Algorithm and approaches to handle large data - A survey," *International Journal of Computer Science and Network (IJCSN),* vol. 2, no. 3, 2013.

[4] K. Bakshi, "Considerations for big data: Architecture and approach," *IEEE Aerospace Conference Proceedings*, 2012.

[5] Global Pulse, "Big data for development: Challenges and opportunities," May 2012.

[6] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. Franklin, S. Shenker, and I. Stoica, "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," in *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*, pp. 2-2, 2012.

[7] A. Bawa-Cavia, "Sensing the Urban: Using location-based social network data in Urban analysis," *Proceedings of the First Workshop on Pervasive Urban Applications*, San Francisco, California, pp. 1-7, 12-15 June 2011.

[8] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, California, pp. 1082-1090, 21-24 August 2011.

[9] L. Tang, and H. Liu, "Leveraging social media networks for classification," *Data Mining and Knowledge Discovery*, vol. 23, no. 3, pp. 447-478, 2011.

[10] F. O. Catak, M. E. Balaban, "CloudSVM: Training an SVM classifier in cloud computing systems," in Q. Zu, B. Hu, and A. Elçi, (eds), Pervasive Computing and the Networked World, ICPCA/SWS 2012, *Lecture Notes in Computer Science*, vol. 7719, Springer, Berlin, Heidelberg, pp. 57-68, 2013.

[11] "The big data and standards market research report," January 2016.

[12] C.-Y. Yeh, W.-P. Su, and S.-J. Lee, "Employing multiple-kernel support vector machines for counterfeit banknote recognition," *Applied Soft Computing*, vol. 11, no. 1, pp. 1439-1447, 2011.

[13] K. I. Kim, K. Jung, S. H. Park, and H. J. Kim, "Support vector machines for texture classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 11, pp. 1542-1550, 2002.

[14] J. Vaidya, B. Shafiq, W. Fan, D. Mehmood, and D. Lorenzi, "A random decision tree framework for privacy-preserving data mining," *IEEE Transactions on Dependable and Secure Computing*, vol. 11, no. 5, pp. 399-411, 2014.

[15] S. Desai, and S. T. Patil, "Differential evolution algorithm with support vector machine to classify objects efficiently," *International Journal of Advance Research in Computer Science and Management Studies (IJARCSMS),* vol. 2, no. 3, pp. 71-74, March 2014.

[16] S. Desai, and S. T. Patil, "Efficient regression algorithms for classification of social media data," 2015 *International Conference on Pervasive Computing (ICPC)*, IEEE, 2015.

[17] S. Desai, A. Fakaria, P. Saini, and S. Sinha, "Analyzing trends in social media marketing," *IJCA*, December 2014.

[18] T. Mitchell, "Decision tree learning," Princeton University.

[19] L. Tang, Z. Ni, H. Xiong, and H. Zhu, "Locating targets through mention in twitter," *World Wide Web*, vol. 18, no. 4, pp. 1019-1049, Springer, 2015.

[20] M. Naaman, J. Boase, and C.-H. Lai, "Is it really about me?: Message content in social awareness streams," *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW'10),* pp. 189-192, 06-10 February 2010.

[21] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," *COLING 2010*, pp. 241-249, ACL, Stroudsburg, 2010.

[22] J. Hannon, M. Bennett, and B. Smyth, "Recommending twitter users to follow using content and collaborative filtering approaches," *Proceedings of the 2010 ACM Conference on Recommender Systems (RecSys'10),* Barcelona, Spain, 26-30 September 2010.

[23] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?," *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*, pp. 591-600, 26-30 April 2010.

[24] S. Burton, and A. Soboleva, "Interactive or reactive? Marketing with twitter," *Journal of Consumer Marketing*, vol. 28, no. 7, pp. 491-499, 2011.

[25] Y. S. Kim, and V. Tran, "Selecting core target users for online social networking marketing with target marketing: A preliminary report," *Proceedings of the Seventeenth Americas Conference on Information Systems*, Detroit, Michigan, 04-07 August 2011.

[26] G. Adomavicius, and Y. Kwon, "Improving aggregate recommendation diversity using ranking-based techniques," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 5, pp. 896-911, 2012.

[27] L. Tang, and H. Liu, "Leveraging social media networks for classification," *Data Mining and Knowledge Discovery*, vol. 23, no. 3, pp. 447-478, 2011.

[28] W. Dai, and W. Ji, "A MapReduce Implementation of C4.5 decision tree algorithm," *International Journal of Database Theory and Application*, vol. 7, no. 1, pp. 49-60, 2014.

[29] I. Frías-Blanco, J. del Campo-Ávila, G. Ramos-Jiménez, R. Morales-Bueno, A. Ortiz-Díaz, and Y. Caballero-Mota,, "Online and non-parametric drift detection methods based on hoeffding's bounds," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 3, pp. 810-823, March 2015.

[30] M. Karim, and M. R. Rahman, "Decision tree and Naïve Bayes algorithm for classification and generation of actionable knowledge for direct marketing," *Journal of Software Engineering and Applications*, vol. 6, no. 4, pp. 196-206, 2013.

[31] P. Zhang, X. Wu, X. Wang, and S. Bi, "Short-term load forecasting based on big data technologies," *CSEE Journal of Power and Energy Systems*, vol. 1, no. 3, pp. 59-67, 2015.

[32] http://www.bestcolleges.com/resources/top-5-mental-health-problems-facing-college-students/

[33] D. Wang, A. Al-Rubaie, A. A. Dhanhani, and J. Ng, "Smart text-classification of user-generated data in educational social networks," *2015 IEEE Frontiers in Education Conference (FIE),* IEEE, El Paso, TX, USA, pp. 1-5, 2015.

[34] S. Cetintas, L. Si, H. P. Aagard, K. Bowen, and M. Cordova-Sanchez, "Microblogging in a classroom: Classifying students' relevant and irrelevant questions in a microblogging-supported classroom," *IEEE Transactions on Learning Technologies*, vol. 4, no. 4, pp. 292-300, October-December 2011.

[35] http://www.technicianonline.com/opinion/article_d1142b70-5a92-11e5-86b4-cb7c98a6e45f.html