ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Analysis of Students Performance Using Data Mining

D. Mahalakshmi[1],

M.Phil, Research Scholar, Department of Computer Science, Jamal Mohamed College, Tiruchirapalli, India

ABSTRACT: Students' informal conversations on social media (e.g. Twitter, Facebook) shed light into their educational experiences opinions, feelings, and concerns about the learning process. Data from such un-instrumented environments can provide valuable knowledge to inform student learning. Analyzing such data, however, can be challenging. The complexity of students' experiences reflected from social media content requires human interpretation. However, the growing scale of data demands automatic data analysis techniques. In this paper, we developed a workflow to integrate both qualitative analysis and large-scale data mining techniques. We focused on engineering students' Twitter posts to understand issues and problems in their educational experiences.

We first conducted a qualitative analysis on samples taken from about 25,000 tweets related to engineering students' college life. We found engineering students encounter problems such as heavy study load, lack of social engagement, and sleep deprivation. Based on these results, we implemented a multi-label classification algorithm to classify tweets reflecting students' problems. We then used the algorithm to train a detector of student problems from about 35,000 tweets streamed at the geo-location of Purdue University. In the proposed work we include two different algorithms such as c4.5 and Support Vector Machine (SVM). C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan.[1] C4.5 is an extension of Quinlan's earlier ID3 algorithm.

The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. This work, for the first time, presents a methodology and results that show how informal social media data can provide insights into students' experiences.

KEYWORDS: SVM, C4.5, ID3, Mining Ontology, Student Performance, Educational Mining.

I. INTRODUCTION

To learn student experience from social media like twitters using workflow. To integrate both qualitative analysis and large-scale data mining techniques. To explore engineering students' informal conversations on Twitter. In order to understand issues and problems students encounter in their learning experiences.

DOI: 10.15680/IJIRCCE.2017. 0503393

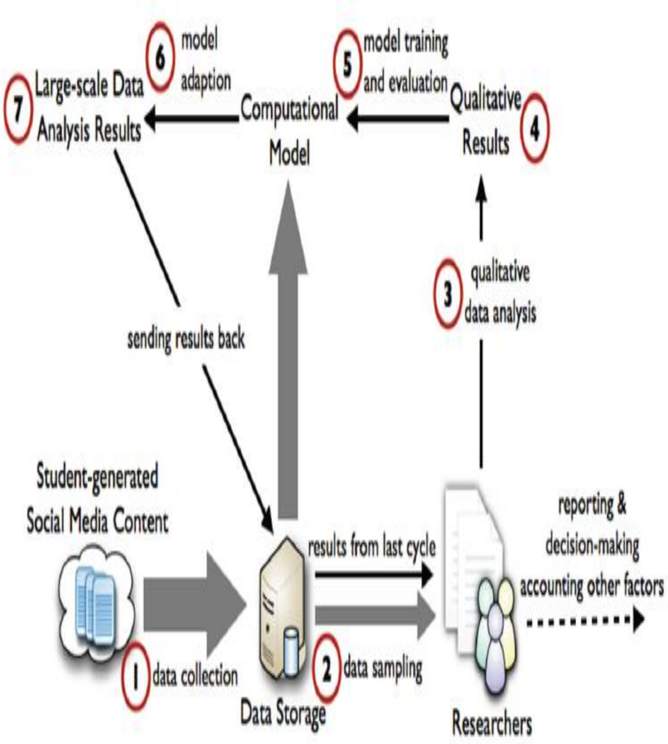


Fig.1. Workflow of Social Media Data

Twitter about problems in their educational experiences they are:- Heavy study load, Lack of social engagement, Negative emotion, Sleeping problems. Learning analytics and educational data mining has focused on analyzing structured data obtained from course management systems (CMS). Classroom technology usage or controlled online learning environments to inform educational decision-making. Twitter is a popular social media site. Its content is mostly public and very concise (no more than 140 characters per tweet). Twitter provides free APIs that can be used to stream data.

Therefore, I chose to start from analyzing students' posts on twitter. Naive Bayesian Classification algorithm to be used in these concepts. I found Naive Bayes classifier to be very effective on dataset compared with other state-of-the-art multi-label classifiers. Social media sites such as Twitter, Facebook, and YouTube provide great venues for students to share joy and struggle, vent emotion and stress, and seek social support. On various social media sites, students discuss and share their everyday encounters in an informal and casual manner.

Students' digital footprints 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 at-risk students, improvement of education quality, and thus enhance student recruitment, retention, and success. The abundance of social media data provides opportunities 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 unpredictability of locations, and timing of students posting on the web, as well as the complexity of students' experiences. 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.

Traditionally, educational researchers have been using methods such as surveys, interviews, focus groups, classroom activities to collect data related to students' learning experiences. These methods are usually very time consuming, thus cannot be duplicated or repeated with high frequency. The scale of such studies is also usually limited. In addition, when prompted about their experiences, students need to reflect 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 systems (CMS), classroom technology usage, or controlled online learning environments to inform educational decision-making. However, to the best of our knowledge, there is no research found to directly mine and analyze student posted content from uncontrolled spaces on the social web with the clear goal of understanding students' learning experiences.

II. PUBLIC DISCOURSE ON THE WEB

The theoretical foundation for the value of informal data on the web can be drawn from Goffman's theory of social performance. Although developed to explain face-toface interactions, Goffman's theory of social performance is widely used to explain mediated interactions on the web today. One of the most fundamental aspects of this theory is the notion of front-stage and back-stage of people's social performances. Compared with the frontstage, the relaxing atmosphere of back-stage usually encourages more spontaneous actions. Whether a social setting is front-stage or back-stage is a relative matter.

For students, compared with formal classroom settings, social media is a relative informal and relaxing back-stage. When students post content on social media sites, they usually post what they think and feel at that moment. In this sense, the data collected from online conversation may be more authentic and unfiltered than responses to formal research prompts. These conversations act as a zeitgeist for students' experiences.

Many studies show that social media users may purposefully manage their online identity to "look better" than in real life. Other studies show that there is a lack of awareness about managing online identity among college students, and that young people usually regard social media as their personal space to hang out with peers outside the sight of parents and teachers.

Students' online conversations reveal aspects of their experiences that are not easily seen in formal classroom settings, thus are usually not documented in educational literature. The abundance of social media data provides opportunities but also presents methodological difficulties for analyzing large-scale informal textual data. The next section reviews popular methods used for analyzing Twitter data.

III. MINING TWITTER DATA

Researchers from diverse fields have analyzed Twitter content to generate specific knowledge for their respective subject domains. For example, Gaffney analyzes tweets with hash tag #iran Election using histograms, user networks, and frequencies of top keywords to quantify online activism. Similar studies have been conducted in other fields including healthcare, marketing, and athletics, just to name a few. Analysis methods used in these studies usually include qualitative content analysis, linguistic analysis, network analysis, and some simplistic methods such as word clouds and histograms.

In our study, we built a classification model based on inductive content analysis. This model was then applied and validated on a brand new dataset. Therefore, we emphasize not only the insights gained from one dataset, but also the application of the classification algorithm to other datasets for detecting student problems. The human effort is thus augmented with large-scale data analysis. Below we briefly review studies on Twitter from the fields of data mining, machine learning, and natural language processing. These studies usually have more emphasis on statistical models and algorithms. They cover a wide range of topics including information propagation and diffusion, popularity prediction, event detection, topic discovery, and tweet classification, to name a few.

IV. LEARNING ANALYTICS AND EDUCATIONAL DATA MINING

Learning analytics and educational data mining (EDM) are data-driven approaches emerging in education. These approaches analyze data generated in educational settings to understand students and their learning environments

DOI: 10.15680/IJIRCCE.2017. 0503393

in order to inform institutional decision-making. The present paper extends the scope of these approaches in the following two aspects. First, data analyzed using these approaches typically are structured data including administrative data (e.g., high school GPA and SAT scores), and student activity and performance data from CMS (Course Management Systems) or VLE (Virtual Learning Environments) such as Blackboard (http://www.blackboard.com/).

IV. INDUCTIVE CONTENT ANALYSIS

Because social media content like tweets contain a large amount of informal language, sarcasm, acronyms, and misspellings, meaning is often ambiguous and subject to human interpretation. Rost et. al argue that in large scale social media data analysis, faulty assumptions are likely to arise if automatic algorithms are used without taking a qualitative look at the data. We concur with this argument, as we found no appropriate unsupervised algorithms could reveal in-depth meanings in our data. For example, LDA (Latent Dirichlet Allocation) is a popular topic modeling algorithm that can detect general topics from very large scale data. LDA has only produced meaningless word groups from our data with a lot of overlapping words across different topics.

There were no pre-defined categories of the data, so we needed to explore what students were saying in the tweets. Thus, we first conducted an inductive content analysis on the #engineering Problems dataset. Inductive content analysis is one popular qualitative research method for manually analyzing text content. Three researchers collaborated on the content analysis process.

C4.5 Algorithm

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan.[1] C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. It became quite popular after ranking #1 in the Top 10 Algorithms in Data Mining pre-eminent paper published by Springer LNCS in 2008 [2]. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set S=S1, S2, S3...of already classified samples. Each sample {\displaystyle s_{i}} s_{i} consists of a p-dimensional vector {\displaystyle (x_{1,i},x_{2,i},...,x_{p,i})} (x_{{1,i}},x_{{2,i}},...,x_{{p,i}}), where the {\displaystyle x_{j}} x_{j} represent attribute values or features of the sample, as well as the class in which {\displaystyle s_{i}} s_{i} falls.

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision.

The C4.5 algorithm then recurs on the smaller sublists. This algorithm has a few base cases. All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class. None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.

Pseudocode

- Check for the above base cases.
- For each attribute a, find the normalized information gain ratio from splitting on a.
- Let a_best be the attribute with the highest normalized information gain.
- Create a decision node that splits on a_best.
- Recur on the sublists obtained by splitting on a_best, and add those nodes as children of node.

Improvements

C4.5 made a number of improvements to ID3. Some of these are:

Handling both continuous and discrete attributes - In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.

Handling training data with missing attribute values - C4.5 allows attribute values to be marked as? for missing. Missing attribute values are simply not used in gain and entropy calculations.

Handling attributes with differing costs.

DOI: 10.15680/IJIRCCE.2017. 0503393

Pruning trees after creation - C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.
Quinlan went on to create C5.0 and See5 (C5.0 for Unix/Linux, See5 for Windows) which he markets commercially. C5.0 offers a number of improvements on C4.5. Some of these are:[5][6]
Speed - C5.0 is significantly faster than C4.5 (several orders of magnitude)
Memory usage - C5.0 is more memory efficient than C4.5
Smaller decision trees - C5.0 gets similar results to C4.5 with considerably smaller decision trees.

Support Vector Machine

In machine learning, support vector machines (SVMs, also support vector networks[1]) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier.

An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The clustering algorithm which provides an improvement to the support vector machines is called support vector clustering and is often used in industrial applications either when data is not labeled or when only some data is labeled as a preprocessing for a classification pass.

VI. LITERATURE SURVEY

In 2011, the authors named G. Siemens and P. Long, introducing a paper called Penetrating the fog: Analytics in learning and education [1], in which it they illustrates the concept of attempts to imagine the future of education often emphasize new technologies ubiquitous computing devices, flexible classroom designs, and innovative visual displays. But the most dramatic factor shaping the future of higher education is something that we can’t actually touch or see: big data and analytics.

Basing decisions on data and evidence seems stunningly obvious, and indeed, research indicates that data-driven decision-making improves organizational output [1] and productivity.1 for many leaders in higher education, however, experience and “gut instinct” [1] have a stronger pull. Meanwhile, the move toward using data and evidence to make decisions is transforming other fields. Notable is the shift from clinical practice to evidence-based medicine in health care. The former relies on individual physicians basing their treatment decisions on their personal experience with earlier patient cases. The latter is about carefully designed data collection that builds up evidence on which clinical decisions are based.

Medicine is looking even further toward computational modeling by using analytics to answer the simple question “who will get sick?” and then acting on those predictions to assist individuals in making lifestyle or health changes [1]. Insurance companies also are turning to predictive modeling to determine high-risk customers. Effective data analysis can produce insight into how lifestyle choices and personal health habits affect long-term risks.4 Business and governments too are jumping on the analytics and data-driven decision-making trends, in the form of “business intelligence.” Higher education, a field that gathers an astonishing array of data about its “customers,” has traditionally been inefficient in its data use, often operating with substantial delays in analyzing readily evident data and feedback [1]. Evaluating student dropouts on an annual basis leaves gaping holes of delayed action and opportunities for intervention.

VII. RESULT ANALYSIS

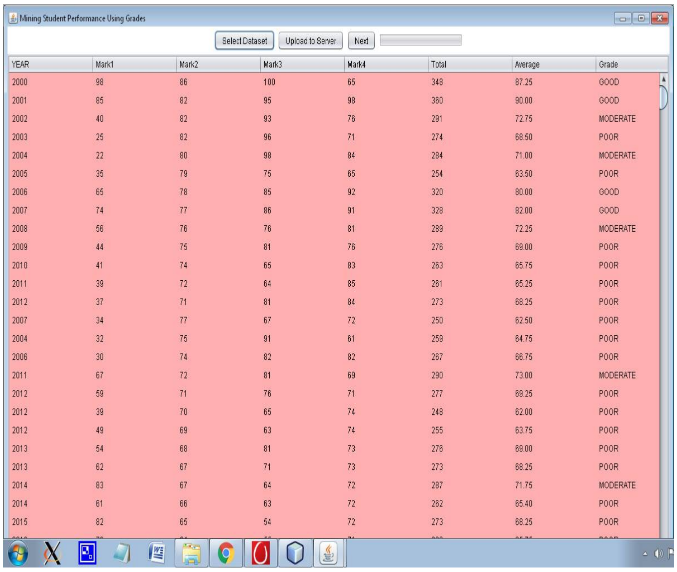
Mastering machine learning algorithms is not a myth at all. Most of the beginners start by learning regression. It is simple to learn and use, but does that solve our purpose? Of course not! Because, you can do so much more than just Regression! Think of machine learning algorithms as an armory packed with axes, sword, blades, bow, dagger etc. You have various tools, but you ought to learn to use them at the right time.

Methodology	Threshold Probability	Accuracy	Precision	Recall	Functional Flow	Average Guessing	Random Guessing
SVM	0	0.1724	0.1724	1.0000	0.2931	0.2941	0.2551
Naïve Bayes Classifier	1	0.0425	0.0054	0.0044	1.5932	1.2541	5.2551
Multi-Label Classification	1	0.2724	0.2724	1.2220	0.3951	0.3941	0.3551
Iterative Algorithm	1	0.0235	0.0064	0.0544	2.4432	2.3341	4.331
Text Mining Appliance	2	0.3724	0.5724	2.2220	1.3951	1.3941	5.3551
Automatic Algorithm	1	0.1235	0.0044	1.0544	0.4432	5.3341	8.331
Time Estimation Algorithm	1	1.3324	0.5724	5.2220	6.5841	5.6641	8.3566
Modelling Algorithm	2	0.0935	0.0764	3.0544	4.689	3.3388	5.5831

As an analogy, think of 'Regression' as a sword capable of slicing and dicing data efficiently, but incapable of dealing with highly complex data. On the contrary, 'Support Vector Machines' is like a sharp knife – it works on smaller datasets, but on them, it can be much stronger and powerful in building models. In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

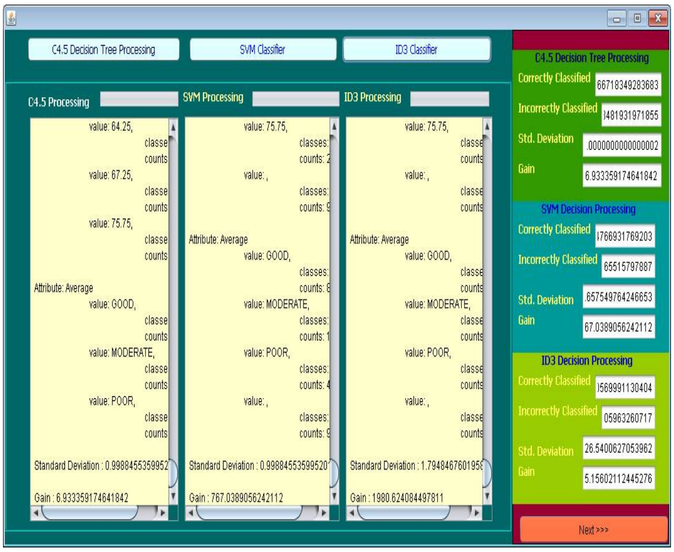
Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

VIII. EXPERIMENTAL RESULTS



YEAR	Mark1	Mark2	Mark3	Mark4	Total	Average	Grade
2000	98	86	100	65	348	87.25	GOOD
2001	85	82	95	98	360	90.00	GOOD
2002	40	82	93	76	291	72.75	MODERATE
2003	25	82	96	71	274	68.50	POOR
2004	22	80	98	84	284	71.00	MODERATE
2005	35	79	75	65	254	63.50	POOR
2006	65	78	85	92	320	80.00	GOOD
2007	74	77	86	91	328	82.00	GOOD
2008	56	76	76	81	289	72.25	MODERATE
2009	44	75	81	76	276	69.00	POOR
2010	41	74	65	83	263	65.75	POOR
2011	39	72	64	85	261	65.25	POOR
2012	37	71	81	84	273	68.25	POOR
2007	34	77	67	72	250	62.50	POOR
2004	32	75	91	61	259	64.75	POOR
2006	30	74	82	82	267	66.75	POOR
2011	67	72	81	69	290	73.00	MODERATE
2012	59	71	76	71	277	69.25	POOR
2012	39	70	65	74	248	62.00	POOR
2012	49	69	63	74	255	63.75	POOR
2013	54	68	81	73	276	69.00	POOR
2013	62	67	71	73	273	68.25	POOR
2014	83	67	64	72	287	71.75	MODERATE
2014	61	66	63	72	262	65.40	POOR
2015	82	65	54	72	273	68.25	POOR

Fig.2. Uploading Dataset into Server



C4.5 Processing

```
            value: 64.25,
                        classe
                        counts
            value: 67.25,
                        classe
                        counts
            value: 75.75,
                        classe
                        counts
Attribute: Average
            value: GOOD,
                        classe
                        counts
            value: MODERATE,
                        classe
                        counts
            value: POOR,
                        classe
                        counts

Standard Deviation : 0.9988455359952

Gain : 6.933359174641842
```

SVM Processing

```
            value: 75.75,
                        classes:
                        counts: 2
            value: ,
                        classes:
                        counts: 9
Attribute: Average
            value: GOOD,
                        classes:
                        counts: 8
            value: MODERATE,
                        classes:
                        counts: 1
            value: POOR,
                        classes:
                        counts: 4
            value: ,
                        classes:
                        counts: 9

Standard Deviation : 0.99884553599520

Gain : 767.0389056242112
```

ID3 Processing

```
            value: 75.75,
                        classe
                        counts
            value: ,
                        classe
                        counts
Attribute: Average
            value: GOOD,
                        classe
                        counts
            value: MODERATE,
                        classe
                        counts
            value: POOR,
                        classe
                        counts
            value: ,
                        classe
                        counts

Standard Deviation : 1.794846760195

Gain : 1980.624084497811
```

C4.5 Decision Tree Processing

Correctly Classified 66718349283683

Incorrectly Classified 3481931971855

Std. Deviation .0000000000000002

Gain 6.933359174641842

SVM Decision Processing

Correctly Classified 1766931769203

Incorrectly Classified 65515797887

Std. Deviation .657549764246653

Gain 67.0389056242112

ID3 Decision Processing

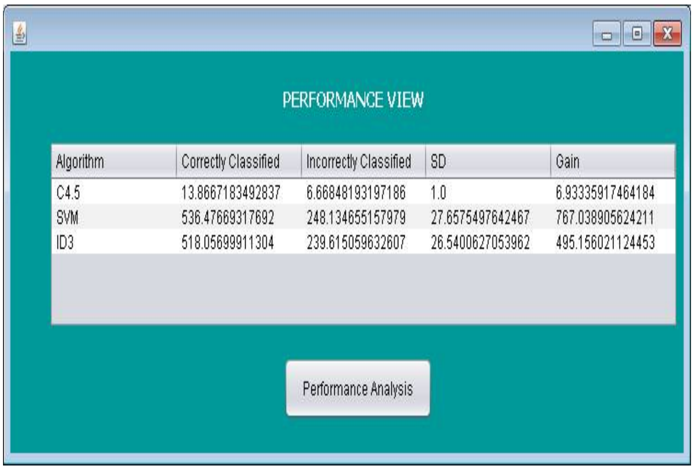
Correctly Classified)569991130404

Incorrectly Classified 05963260717

Std. Deviation 26.5400627053962

Gain 5.15602112445276

Fig.3. Classification Process



PERFORMANCE VIEW

Algorithm	Correctly Classified	Incorrectly Classified	SD	Gain
C4.5	13.8667183492837	6.66848193197186	1.0	6.93335917464184
SVM	536.47669317692	248.134655157979	27.6575497642467	767.038905624211
ID3	518.05699911304	239.615059632607	26.5400627053962	495.156021124453

Fig.4. Performance Overview

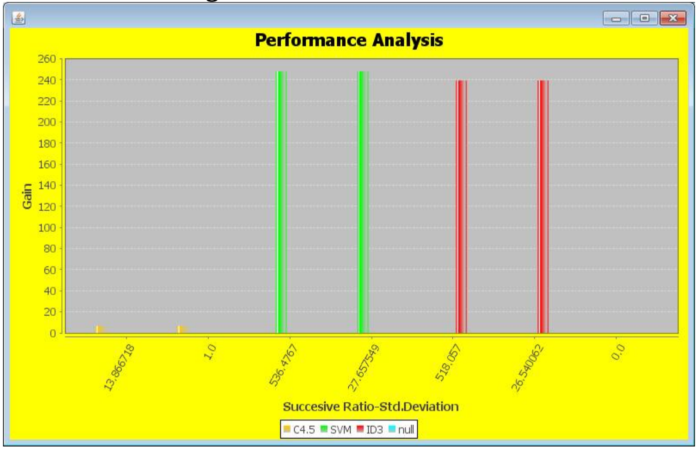


Fig.5. Performance Anlaysis

DOI: 10.15680/IJIRCCE.2017. 0503393

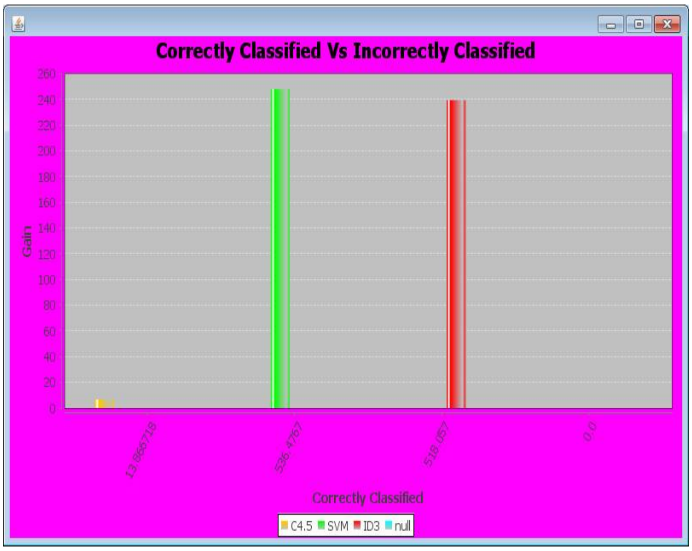


Fig.6. Classification Analsysis

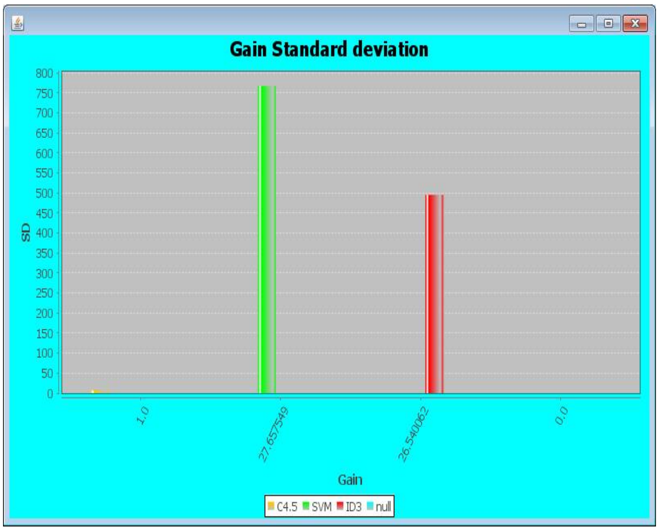


Fig.7. Standard Deviation and Gain

IX. CONCLUSION

Learning Analytics is an emerging new research field with many tools, which offer many valuable services to educators for monitoring and tracking learners' interactions in online learning environments. This research is to design and implement application to track students' online learning activities based on LMS logs and semantic similarity between sentences in messages post based on semantic and word order information. Semantic similarity is derived from a lexical knowledge base and a corpus. The lexical knowledge base models common human knowledge about words in a natural language this knowledge is usually stable across a wide range of language application areas A corpus reflects the actual usage of language and words. LMS and made some small interface changes to Moodle. Intuitively, the presence of the system popularity bars should encourage students to check the course materials more frequently and promptly if he or she sees most of their classmates have already done so. Our hypothesis is that the availability of system statistics will positively affect both how an instructor adapts the course and how students learn.

Monitoring student learning activity is an essential component of high quality education, and is one of the major predictors of effective teaching (Cotton, 1998). Research has shown that letting a learner know his or her ranking in the group results in more effective learning. This application provides a possible means for students and the instructor to receive this feedback. Future work will build a smart student profile from different activities on LMS (Excellent, V-Good, Good, Poor). The messages will Classify in online discussion semantic, by the messages categories (Seminar, Question, Argumentation, Counter- Argumentation, Others).

REFERENCES

[1] Macfadyen, L. P., & Dawson, S. Numbers Are Not Enough. Why e-Learning Analytics Failed to Inform an Institutional Strategic Plan. Educational Technology & Society, 15 (3), 149–163, 2012.

[2] G. Lazakidou, and S. Retalis, “Using computer supported collaborative learning strategies for helping students acquire self-regulated problemsolving skills in mathematics”, Computers & Education, vol. 54(1), pp.3-13, 2010.

[3] Sujana Jyothi, Claire McAvinia*, John Keating ."A visualisation tool to aid exploration of students’ interactions in asynchronous online communication" Computers & Education, Volume 58, Issue 1, January,Pages30-42 , 2012

[4] R. Mazza, and L. Botturi, “Monitoring an online course with the GISMO tool: A case study”, Journal of Interactive Learning Research, vol.18(2), pp.251-265, 2007.

[5] Riccardo Mazza, Marco Bettoni, Marco Faré, Luca Mazzola. MOCLog – Monitoring Online Courses with log data. In: In Retalis, S., and Dougiamas, M."1st Moodle Research Conference Proceedings", pp. 14-15, 2012.

[6] B. Jelen, and M. Alexander, M. “Pivot Table Data Crunching: Microsoft Excel 2010”, Que Corporation, 2010.

[7] F. C. Sampayo, (2013, April 22), Analytics and Recommendations. Available:https://moodle.org/plugins/view.php?plugin=block_analytics_recommendations

[8] R. Rada, H. Mili, E. Bichnell, and M. Blettner, “Development and Application of a Metric on Semantic Nets,” IEEE Trans. System, Man, and Cybernetics, vol. 9, no. 1, pp. 17-30, 1989.

[9] Resnik, Philip. Using information content to evaluate semantic similarity. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, pages 448–453,1995

[10] Lin, Dekang. Automatic retrieval and clustering of similar words. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING–ACL ’98), Montreal, Canada, pages 768–774, ,1998.

DOI: 10.15680/IJIRCCE.2017. 0503393