DOI: http://dx.doi.org/10.26483/ijarcs.v8i9.4899

ISSN No. 0976-5697

Volume 8, No. 9, November-December 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

-ANALYSIS OF RELEVANCY USING RAPID MINER WITH AN EXPERIMENTAL STUDY ON POPULAR DISCUSSION FORUM

P.Radhika
Research Scholar
Rayalaseema University
Kurnool

Dr P.SureshVarma
Professor of CSE
Adikavi Nannaya University
Rajahmundry

M.Santosh
Student
VNR VJIET
Hyderabad

Dr.G.V.Rao
CEO, SasankaELearning Pvt Ltd,
Hyderabad

K.Kamakshaiah
Asst .Professor, Dept of CSE
VNR VJIETHyderabad

Abstract: Massive Open Online Courses (MOOCs) is a model for delivering course content online independent of attendance and place .Online Discussion not only brings opportunities for innovation in education but also shifts focus from traditional educations. MOOCs in Computer Programming can help in attracting students because it personalizing their learning experiences at lower cost. In addition to watching videos, students also engage in discussion forums to share their understanding. On the other hand, online discussion forums provide a platform to ingenious students to share their ideas in a holistic way which is not possible through regular websites, videos, online courses. Both novice and expert users search the web exhaustively for their coding practises to learn gradations behind libraries, programming languages and frameworks.In community-based question-answering communities, where students ask questions there is no guarantee that students get their what they are searching for. This poses an unsatisfactory level in the student. In this paper, we present a machine learning model that predicts the relevancy of answers to the forum questions using historical forum data .The study attempted to identify the relevance criteria that people use when browsing a discussion forum

Keywords: MOOC's, Engagement, Discussion forums, Prediction, Relevancy.

1. INTRODUCTION

MOOCs have become a popular and significant source of distance-education due to the flexibility of course access and varied topics of courses. Online courses provide a good platform for learning across various places with these benefits ,there are million students who enroll in one or more MOOCs. Since MOOCs typically offer with only video-based lectures with no interactions with professor, students usually lack the opportunity to have interactions with other students and the instructor relative to traditional resident courses.

Anderson [3] discusses the features of successful forums , as well as the capacities and boundaries for online learning. Social network analysis is commonly applied to investigate the interactions on learning forums[4]. Dropout behaviour is identified as related to sentiment of students' forum posts based on sentiment analysis of MOOC forums discussions[5].Hauang et al.[6] studied the behaviour of users who make great contributions across different MOOC forums.

Technology-based interactions, such as blogs and forums ,are introduced as potential solutions . The Q&A forum like StackOverflow, Stack exchange, quora are the most popular forums [7], [8], but a lot of niche forums exist for every feasible piece of programming technology. In addition, online documentation pages, technical blog posts, and open-source code repository websites often embed discussion forums at the bottom of each webpage to allow people to discuss that page's contents.

Research using MOOC Forum

Regardless of the prevalent acceptance of discussion forums, the user interface of discussion forums has not changed much in the past years since early avatars such as Usenet newsgroups [9] and The WELL [10]. Most of the discussion forums are comprised of tree of text-based threads and all the other kinds of programming forums look almost identical. Some popular sites such as StackOverflow have incorporated features such as voting, answers ,reputation metrics, category tags, hot networking questions, searching, sorting, filtering, favorites, rich text formatting, active, oldest tags, recent tags and syntax highlighting for code. But in principal, most of them follow generic hierarchy of word-based discussions. When it comes to relevancy of answer to the question asked in discussion forum, most of them are accepted through upvotes. The answer with highest upvotes is treated as the most acceptable answer. However students cannot find the answer what they are searching for and this is one the reason for their drop out ratio from online courses.In this study we attempted to find the relevancy of an answer to the question asked in a popular discussion forum, stack exchange . The data from discussion form is taken and 1000 questions from programming on C is taken and a dataset is prepared out of it. We have applied supervised classification algorithm "Decision Tree" on our dataset using Rapid Miner tool.

Prediction Model

The data provided by the discussion forum is used to define various features. By these features, we can provide many inclusive methods to get the predictions right. The data-set formation is based on the binary classification that is to provide the relevant discussion. These features help for the formation of the data-set that help to predict the relevancy. Features include user features and the thread features.

User features:

As on all the discussion forums, the candidates participate in answering different set of questions which help them to understand or clear the doubts of others. These answering to the different set of questions given in a thread that ultimately have an answer accepted or it may be opinion based which hence can be used for training model. These feature also provide further information about the last activity in the forum and the number of viewers visited that forum . In the formation of the data-set the number of accepted answers are considered to provide better and more accurate predictions.

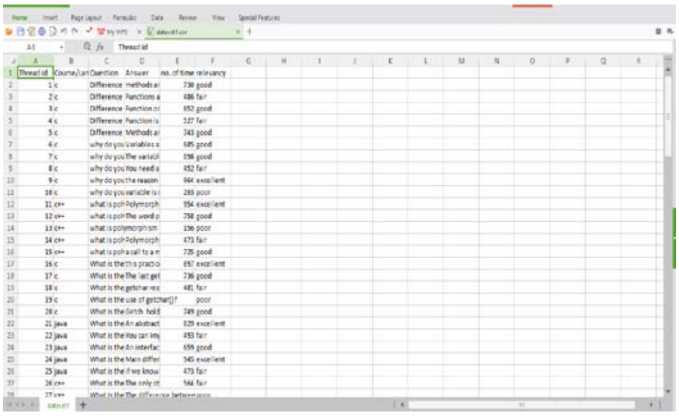
Thread features:

It holds the information about the built-in thread post. Most of the thread post provide information about the answers and the comments that have been already posted.

The questions and the answers are taken from the most of the computer science discussion forums wherein the discussion in these forums are more frequent and the discussion goes on a related pace. The impact of these features are provide the information gain by the number of accepted answers posted by the users, sum of the votes received by an user overall threads ,and number of posted comments by users.

Each extracted feature is then represented into set of distinct values to facilitate better machine learning . For our predication purpose we used RAPID-MINER tool-kit without any other techniques involved to show that our features are sufficient enough to create meaningful predictions.

The programming languages like c is considered for the formation of the formation of the data-set. These languages are mostly discussed languages when compared to the other as they are the base roots for advanced computer programming languages to learn.



	A	B	C	D	E	F
1	Thread id	Course/Lan	Question	Answer	no. of time	relevancy
2	1	c	Difference	methods ar	730	good
3	2	c	Difference	Functions a	486	fair
4	3	c	Difference	Function or	652	good
5	4	c	Difference	Function is	527	fair
6	5	c	Difference	Methods ar	743	good
7	6	c	why do you	Variables a	685	good
8	7	c	why do you	The variabl	698	good
9	8	c	why do you	You need a	452	fair
10	9	c	why do you	the reason	964	excellent
11	10	c	why do you	variable is	283	poor
12	11	c++	what is pol	Polymorph	954	excellent
13	12	c++	what is pol	The word p	738	good
14	13	c++	what is polymorphism		156	poor
15	14	c++	what is pol	Polymorph	473	fair
16	15	c++	what is pol	a call to a m	725	good
17	16	c	What is the	this practio	857	excellent
18	17	c	What is the	The last get	736	good
19	18	c	What is the	getchar rea	481	fair
20	19	c	What is the use of getchar()?			poor
21	20	c	What is the	Ganch hold	549	good
22	21	java	What is the	An abstract	829	excellent
23	22	java	What is the	You can im	493	fair
24	23	java	What is the	An interfac	659	good
25	24	java	What is the	Main differ	945	excellent
26	25	java	What is the	if we know	473	fair
27	26	c++	What is the	The only ot	364	fair
28	27	c++	What is the	The difference between		poor

Fig :1 Sample formation of data-set to provide the relevancy.

We used a classification algorithm called "Decision Tree" which supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.

2. EVALUATION

We took a whole set of 1000 questions from different programming languages and we browsed through the discussion forum to our evaluation we got with a total of 950 questions where answered and the rest of 350 questions where unanswered .Approximately about three-fourth of those questions were responded with at-least two to three answers and the rest with the single answers. Since we used historical data for the training of the model , we made sure to use only the the time stamped features and thus only the data was available when each answer was posted.

Evaluation measures

In a machine learning model, the prediction is based on the formation of the data-set . In the data-set the class label or the attribute (number of times accepted) that is to be used plays a vital role . The algorithms are used to provide the a class probability I.e., for the training data-set probabilities are calculated by using both the algorithms .After the comparison of the algorithms ,these calculations estimates the relevancy in the discussion forum. The analysis can provide which discussion forum is best to refer and discuss

The attributes like questions ,answers , number of times accepted and relevancy are used .However , the number of times accepted is formally divided and a grading format is assigned to make the better and accurate predictions. The customized grading system was based on the number of times the answer is accepted in the discussion forum and also the up-votes to the answer. The grading system includes four categories they are :

Excellent : the range or the accepted points should be between 750-1000

Good : the range or the accepted points should be between 500-749.

Fair : the fair points are between 230-499

Poor : The range or the accepted points should be between 0-229.

We have use this accepted points to give a proper definition rather than manually giving the grading format tag e.g.,if their is a thread for that class label can be placed manually as it is a implementation based we used the number of times the answer is accepted number suppose the answer is accepted in the range of 750 to 1000 then it is given the grade of " Excellent ". Similarly we used this formatting process for the grading in the other attribute's class label section.

As training and evaluation was performed on historical data, we implemented ten fold cross validation to mitigate the impact of over-fitting . We can also use the percentages but the accuracy using it is less when compared to that of cross validation.

We used many features to get the more accurate predictions while training the model .They are:

Number of times accepted : As we discussed above , the answers of the user that had the most answers accepted is made into grade format and given the specific tag I.e., excellent , good, fair and poor.

Most votes received and most comments : The answers of the users from the discussion forums that have the most up-votes across all the forums and also the most comments respectively is considered.

Further , we trained machine learning models on all the features using "Decision Tree Algorithm" (a branching method to illustrate every possible outcome of a decision)

3. EXPERIMENTAL SETUP

We implemented the whole set-up in the Rapid-miner . Rapid-miner is required to be uploaded with a data-set for further operations .The data-set is manually prepared by using some specific formatting process. The available data-set can be be saved as a CSV file / JSON file or excel file. For our proposed experiment we used a CSV file as shown in Fig.

The Rapid-miner provides different platforms to execute your model like Market Based Analysis ,Out-lier detection, Medical fraud detection, Web Analytics etc. We will use a blank process for this experiment and create process model or the design by putting different operators in Fig .2 like retrieve data-set, cross validation rule etc.

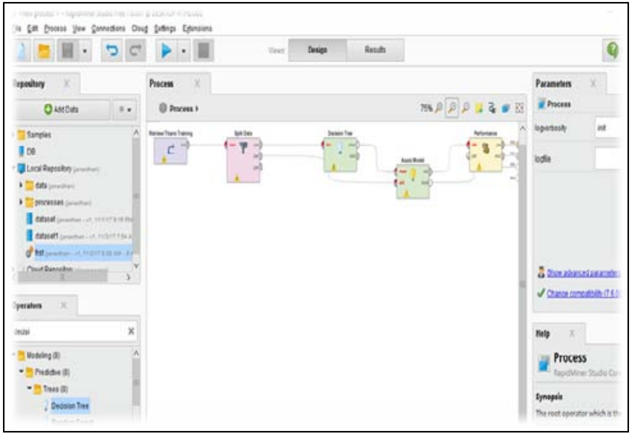


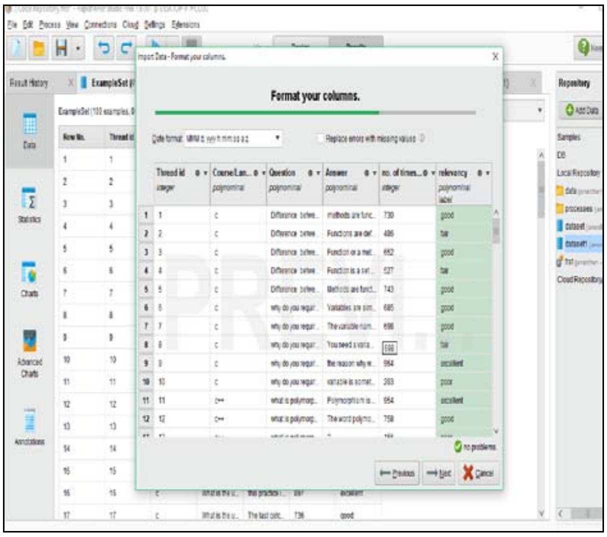
Fig.2 Rapid miner with different operators.

Methodology and Result:

The aim of designing in Rapid-miner is to supply the input along with the validation rule to the decision tree operator and see the results. This set-up give the performance evaluation of decision tree classifier. Following steps are to be in followed in Rapid-miner for setting up the process and getting desired result using decision tree classifier.

1. Create new Blank process in Rapid-miner.
2. Load data-set by using Rapid-miner.

The Rapid miner have different steps while loading the data-set . The data-set should contain a class label and that can be done using change role option that is provided in setting part below the each attribute part .Figure 3 is the loaded data-set into the Rapid-miner for operation.



	Thread id (integer)	Course Lan... (polynominal)	Question (polynominal)	Answer (polynominal)	no. of times... (integer)	relevancy (polynominal label)
1	1	c	Difference betwe...	methods are func...	730	good
2	2	c	Difference betwe...	Functions are def...	486	fair
3	3	c	Difference betwe...	Function or a met...	652	good
4	4	c	Difference betwe...	Function is a set...	527	fair
5	5	c	Difference betwe...	Methods are funct...	743	good
6	6	c	why do you requir...	Variables are sim...	685	good
7	7	c	why do you requir...	The variable nam...	698	good
8	8	c	why do you requir...	You need a varia...	598	fair
9	9	c	why do you requir...	The reason why w...	964	excellent
10	10	c	why do you requir...	variable is somet...	283	poor
11	11	c++	what is polymorp...	Polymorphism is...	954	excellent
12	12	c++	what is polymorp...	The word polymo...	758	good

Fig 3: Loaded data-set into Rapid-miner.

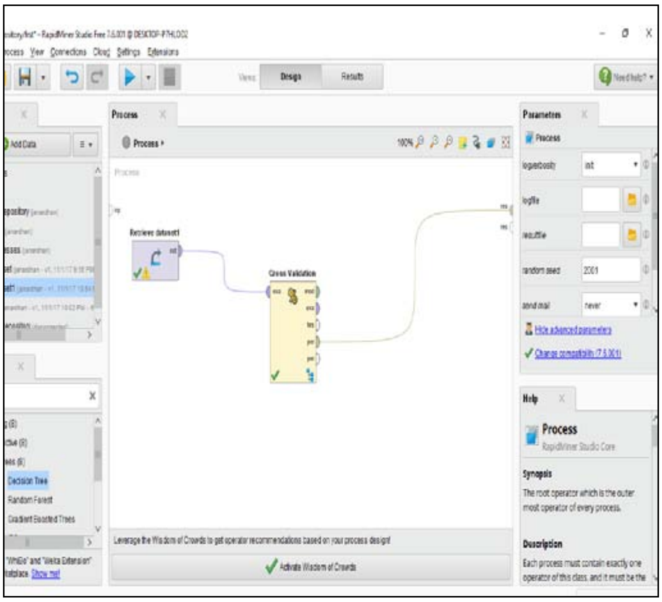
The loaded data is available under 'Local Repository in Rapid-miner.For plotting of data. A new blank process is added I Rapid-miner.

3. Save the process with the respective file name and
4. Run.
5. When the data-set gets loaded in the Rapid-miner provides some different statistical graphs.

For the proposed experiment we required to design a running model which is done in the designing section of the Rapid-miner. Following steps are taken in designing the process model.

1) Retrieves the MOOC data-set from the Repository.

2)Cross validation is connected to the data-set and it splits into two subsets. Stratified sampling is used to retain the distribution of label in both training and test stages.j



logverbosity init
logfile
resultfile
random seed 2001
send mail never
Hide advanced parameters
Change compatibility (7.6.001)

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Activate Wisdom of Crowds

The root operator which is the outer most operator of every process.

Each process must contain exactly one operator of this class, and it must be the

Fig 4: Process model When we double click on the cross validation we get stages like

A) Training stage and

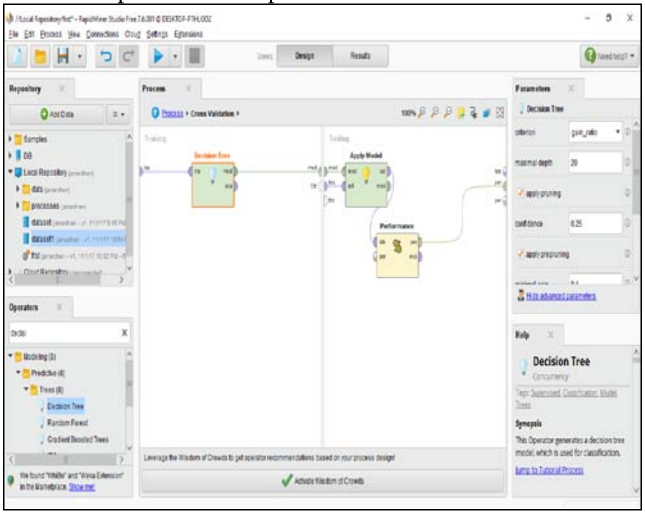
B) Testing stage.

Training stage :

In this stage , we have to include retrieved data-set model connect it to the mod part of the training section . In the operator the classification algorithm that is decision tree is selected and dragged on to the connected link (from retrieve data-set and mod part).The connection is established after this process.

Testing stage:

This stage is very much important to get the statistical measures where in apply model and performance models are connected and that link is given a final link to the res port in the testing stage. The input for this is taken form the connected port that mod port .



Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

This Operator generates a decision tree model, which is used for classification.

Fig 4: Cross Validation design.

3) Finally, go back to the process model, and make a final connection from cross validation to the res port .

4) Hit on run button.

The statistical measures are provided after the whole process have run successfully.

When the process is run, then the following output is generated in following:

I. Simple Distribution

II. Example-set(Apply-Model)

III. Performance Vector.

The above three categories have further classification which are

I Simple Distribution

a)Description

b)Charts(Fig.6)

c)Distribution Table

Ii)Example-set(Apply-Model)

a)Data(Fig.7)

b)Statistics(Fig.8)

c)Charts

d)Advanced Charts

III) Performance Vector

a) %performance

b) Description(Fig.9)

c) Annotations.

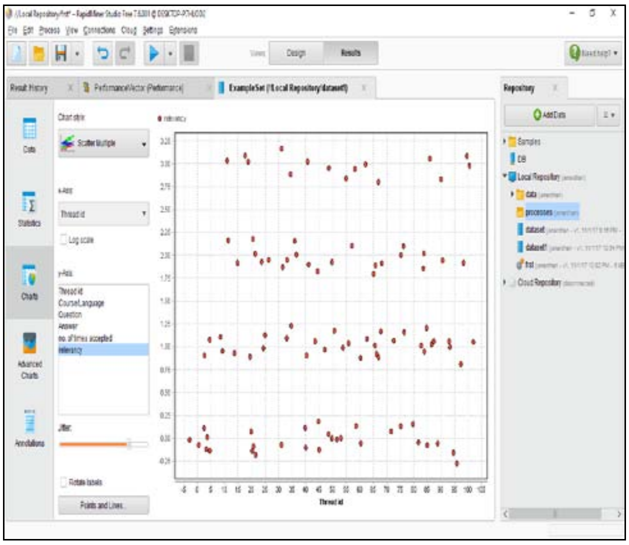
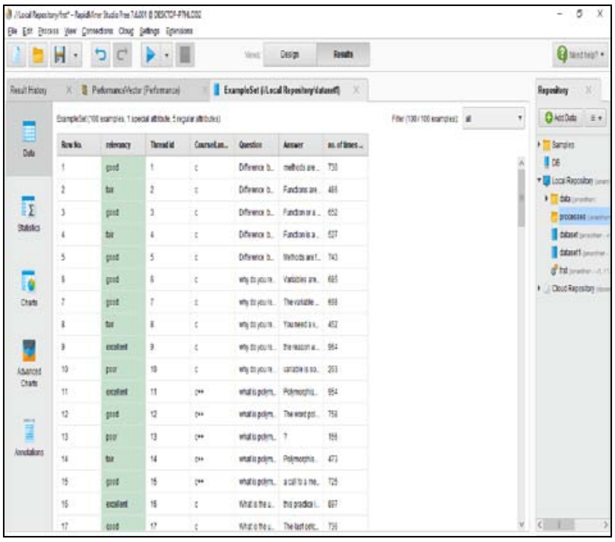


Fig.6: Simple Distribution (Charts)

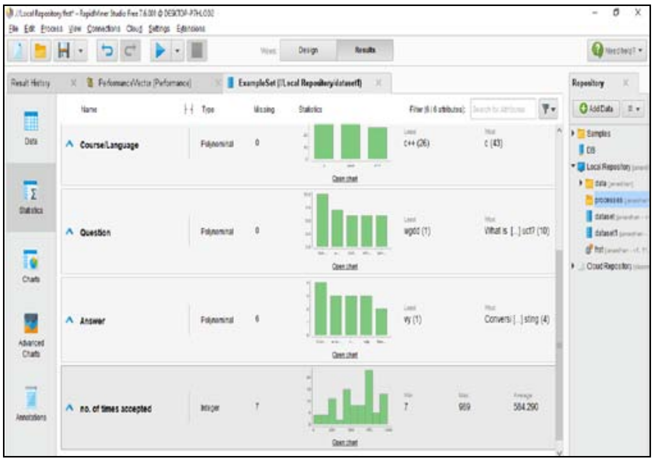

File Edit Process View Connections Cloud Settings Extensions
Design
Results
Result History
PerformanceVector (Performance)
ExampleSet (//Local Repository/dataset)
Data
Statistics
Charts
Advanced Charts
Annotations
Repository
Add Data

ExampleSet (100 examples, 1 special attribute, 5 regular attributes)

Filter (100 / 100 examples): all

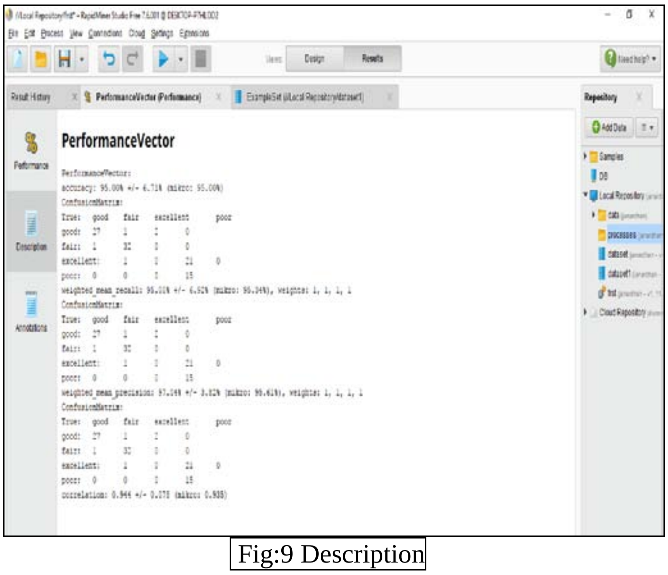
Row No.	relevancy	Thread id	Course/Lan...	Question	Answer	no. of times ...
1	good	1	c	Difference b...	methods are ...	730
2	fair	2	c	Difference b...	Functions are ...	488
3	good	3	c	Difference b...	Function or a ...	652
4	fair	4	c	Difference b...	Function is a ...	537
5	good	5	c	Difference b...	Methods are t...	743
6	good	6	c	why do you re...	Variables are...	685
7	good	7	c	why do you re...	The variable ...	698
8	fair	8	c	why do you re...	You need a s...	452
9	excellent	9	c	why do you re...	the reason a...	984
10	poor	10	c	why do you re...	variable is so...	263
11	excellent	11	c++	what is polym...	Polymorphis...	954
12	good	12	c++	what is polym...	The word pol...	758
13	poor	13	c++	what is polym...	?	156
14	fair	14	c++	what is polym...	Polymorphis...	471
15	good	15	c++	what is polym...	a call to a me...	725
16	excellent	16	c	What is the u...	this practice i...	897
17	good	17	c	What is the u...	The last ...	736

Fig.7: Apply model(data)



Name	Type	Missing	Statistics			
Course/Language	Polynominal	0		Least c++ (26)	Most c (43)	
Question	Polynominal	0		Least wgdd (1)	Most What is [...] uct? (10)	
Answer	Polynominal	6		Least vy (1)	Most Conversi [...] sting (4)	
no. of times accepted	Integer	7		Min 7	Max 989	Average 584.290

fiq 8: statistics


//Local Repository/first* – RapidMiner Studio Free 7.6.001 @ DESKTOP-F74L0O2
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Need help?
Result History
PerformanceVector (Performance)
ExampleSet (//Local Repository/dataset1)
Performance
Description
Annotations
Repository
Add Data
Samples
DB
Local Repository
data
processes
dataset
dataset1
first
Cloud Repository

PerformanceVector

```
PerformanceVector:
accuracy: 95.00% +/- 6.71% (mikro: 95.00%)
ConfusionMatrix:
True:   good    fair    excellent       poor
good:   27      1       [illegible]       0
fair:   1       32      [illegible]       0
excellent:      1       [illegible]       21      0
poor:   0       0       [illegible]       15
weighted_mean_recall: 95.33% +/- 6.52% (mikro: 95.34%), weights: 1, 1, 1, 1
ConfusionMatrix:
True:   good    fair    excellent       poor
good:   27      1       [illegible]       0
fair:   1       32      [illegible]       0
excellent:      1       [illegible]       21      0
poor:   0       0       [illegible]       15
weighted_mean_precision: 97.14% +/- 3.82% (mikro: 95.61%), weights: 1, 1, 1, 1
ConfusionMatrix:
True:   good    fair    excellent       poor
good:   27      1       [illegible]       0
fair:   1       32      [illegible]       0
excellent:      1       [illegible]       21      0
poor:   0       0       [illegible]       15
correlation: 0.944 +/- 0.078 (mikro: 0.938)
```

Fig:9 Description

4. CONCLUSION

Discussion forums offer a great platform for students to share their understanding of the subject. However, when it comes to relevancy of answers posted is relatively poor and some are left unanswered. This increases the dropout ratio of the students consistently. Our work presented here shows the analysis of relevancy of the posts to programming subjects of computer science. The relevancy obtained shows fair grade scale that indicates the discussion in the forum is average. This clearly states that by the formation of proper data-set using different discussion forum data we can find the relevancy also.

REFERENCES

[1] J. Brandt, P. J. Guo, J. Lewenstein, M. Dontcheva, and S. R. Klemmer, “Two studies of opportunistic programming: Interleaving web foraging, learning, and writing code,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ser. CHI ’09. New York, NY, USA: ACM, 2009, pp. 1589–1598.

[2] J. Brandt, M. Dontcheva, M. Weskamp, and S. R. Klemmer, “Examplecentric programming: Integrating web search into the development environment,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ser. CHI ’10. New York, NY, USA: ACM, 2010, pp. 513–522.

[3] Smith, M.K., Wood, W.B., Adams, W.K., Wieman, C., Knight, J.K., Guild, N., Su, T.T.: Why peer discussion improves student performance on in-class concept questions. Science 323, 122–124 (2009)

[4] Stephens-Martinez, K., Hearst, M.A., Fox, A.: Monitoring moocs: which information sources do instructors value? In: Proceedings of the First ACM Conference on Learning@ Scale Conference, pp. 79–88. ACM (2014)

[5] Andresen, M.A.: Asynchronous discussion forums: Success factors, outcomes, assessments, and limitations. Educational Technology & Society 12, 249–257 (2009)

[6] Zhu, E.: Interaction and cognitive engagement: An analysis of four asynchronous online discussions. Instructional Science 34, 451–480 (2006)

[7] Wen, M., Yang, D., Rosè, C.P.: Sentiment analysis in MOOC discussion forums: What does it tell us? In: Proceedings of Educational Data Mining (2014)

[8] Huang, J., Dasgupta, A., Ghosh, A., Manning, J., Sanders, M.: Superposter behavior in mooc forums. In: Proceedings of the First ACM Conference on Learning@ Scale Conference, pp. 117–126. ACM (2014)

[9] Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, “Discovering value from community activity on focused question answering sites: A case study of stack overflow,” in Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD ’12. New York, NY, USA: ACM, 2012, pp. 850–858.

[10] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann, “Design lessons from the fastest q&a site in the west,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ser. CHI ’11. New York, NY, USA: ACM, 2011, pp. 2857–2866.

[11] S. L. Emerson, “Usenet: A Bulletin Board for Unix Users,” Byte magazine, vol. 8, no. 10, pp. 219–236, October 1983.

[12] H. Rheingold, The virtual community : homesteading on the electronic frontier. Reading, Massachusetts: Addison Wesley, 1993.

[13]“edX course: Introduction to Computer Science and Programming Using Python,” https://www.edx.org/course/ introduction-computer-science-mitx-6-00-1x-0, accessed: March 2015.