Classifying Student's Learning Experience using Improved Apriori and CART

Pooja Verma
M.Tech scholar
Department of Computer Science and Engineering, T.I.T (Excellence) Bhopal

Rajesh Boghey
Professor & HOD
Department of Computer Science and Engineering, T.I.T (Excellence) Bhopal

Sandeep Rai
Assistant Professor
Department of Computer Science and Engineering, T.I.T (Excellence) Bhopal

ABSTRACT
Here in this paper a new of classifying Student's learning experience on online social networks such as facebook, twitter is proposed which helps to find various issues and problems in their educational experiences. The existing technique implemented for the classification for the Student's learning experience provides multi-label classification to reflect various problems but fails to provide the improvement in accuracy, hence a new multi-label classification using improved Apriori algorithm is proposed which generates a set of candidate rules and finally classify Student's experience using Classification & Regression Tree. The proposed methodology implemented provides better results in comparison with an existing technique. The experimental results are performed and tested on various parameters such as precision and recall and final Score. The various student's learning experience and their classification is done here using Fuzzy-Apriori and CART provide and better way to final and issue problems in various fields.

Keywords
Online Social Network, Apriori algorithm, Fuzzy rules, Classification & Regression Tree, Decision Tree.

1. INTRODUCTION
Large scale data has dominated every aspect of computing over the past few years and will continue to do so with an ever increasing trend. Big Data applications come in all shapes and sizes where most of the commercially driven use cases tend to have relatively less complex applications consuming colossal amounts of data compared to highly complex applications. A huge challenge with forming a benchmark for big data structures is the extensive collection of difficulty that needs big data explanations. Some of the most frequent applications are a scientific study, intelligence, social media, healthcare, marketing, finance, and retail. It is an open question whether a single benchmark can be produced that is cooperative to all of these areas or whether unusual groups will need differently calculates of performance.As there are many promising big data purposes, they acquire an incremental and iterative approach as an alternative to a top-down approach. Initially, they examine the leading application domains of internet services—a significant class of big data applications according to extensively suitable metrics—the number of page views and on a daily basis visitors [1].

According to the definition of Big Data, Big Data is characterized by volume, velocity, and variety where traditional data processing methods and tools cannot be qualified. Volume means a very large amount of data, particularly in data storage and computation. By 2010 the global amount of information would rapidly up to 988 billion GB [2]. Experts predict that by 2020 annual data will increase 43 times. Velocity means the speed of data growth is increasing, meanwhile, people's requirements for data storage and processing speed are also rising. Purely in scientific research, the annual volume of new data accumulated by the Large Hadron Collider is about 15PB [3]. In the field of electronic commerce, Wal-Mart's sells every day more than 267 million (267Million) products [4]. Data processing requires faster speed, and in many areas data have been requested to carry out in real-time processing such as disaster prediction and rapid disaster rehabilitation under certain conditions need quickly quantify the extent of the disaster, the regional scope impacted and etc. Variety refers to the data that contains structured data table, semi-structured and unstructured text, video, images and other information, and the interaction between data is very frequent and widespread. It specifically includes diverse data sources, various data types, and a strong correlation between the data.

Figure-1: Big Data Application Domains.

With the development of computer and network technology, as well as intelligent systems is commonly used in modern life, big data has become increasingly close to people's daily lives. In 2008, Big Data issue released by "Nature" pointed out the importance of big data in biology, and it was necessary to build a biological big data system to solve complex biological data structure problem [5]. Paper [5] pointed out that the new big data system must be able to tolerate various structures of data and unstructured data, has flexible operability and must ensure data reusability. Furthermore, Big Data plays an important role in the defense of national network digital security, maintaining social stability and promoting sustainable economic and social development [6].

As the big data industry persists to produce and start widespread requires and developments, significant benchmarks will be a method to evaluate different schemes and permit engineers to plan better explanations and consumers to make knowledgeable acquires. There have been

a number of efforts at creating big data benchmarks [7-9]. None of them has increased extensive recognition and large procedure. It continues an indefinable objective to estimate an extensive range of projected big data solutions. The area of big data performance is in a condition where every learning and maintain utilizes a different method. Results from one publication to the subsequently are not equivalent and frequently not even intimately associated, as it was the case for OLTP some twenty years ago and for choice sustain abruptly subsequently. We distinguish the need for a measure to determine the performance of big data schemes. The report market research and deal publications reporting of the subject point outs that big data is quick distribution within the commercial IT communications, even for non-technology, conventional industry areas. This novel phase in the expansion of big data presents a chance to be familiar with considerable application domains as they appear, so as to describe appropriate and intention benchmarks.

We dispute that the area as an entire has not increased enough knowledge to authoritatively pronounce "the big data benchmark should be X." Big data keep on a novel and fast altering area. The intrinsic complication, variety, and extent of such schemes initiate extra challenges for essential a representative, convenient and scalable benchmark. On the other hand, by uniting earlier period knowledge from TPC-method benchmarks and promising to approach from MapReduce uses they can as a minimum illuminate some key apprehensions and perceptions related to building usual big data benchmarks.Big data applications, but infrequently use again input data and this policy for data demanding applications do not effort in many cases. The modern computational situation has been and is developing essentially for accelerating of benchmarks i.e. LINPACK or SPEC. These benchmarks are comparatively scalable according to a number of CPUs. Big data applications are not scalable to the different and the existing computational situation is not essentially perfect for big data applications.Big data benchmarks are the establishment of those attempts [10]. On the other hand, the complication, variety, frequently altered workloads—so called workload mix [11] and quick development of big data systems inflict enormous challenges to big data benchmarking.

Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

Apriori

The Apriori Algorithm is an influential algorithm for mining frequent itemsets for boolean association rules. • Apriori uses a "bottom-up" approach, where frequent subsets are extended one item at a time (a step is known as candidate generation, and groups of candidates are tested against the data.

Fuzzy Rules

A fuzzy rule is defined as a conditional statement in the form: IF x is A. THEN y is B. where x and y are linguistic variables; A and B are linguistic values determined by fuzzy sets on the universe of discourse X and Y, respectively.

Online Social Networks

A social networking service is an online service, platform, or site that focuses on facilitating the building of social networks or social relations among people who, for example, share interests, activities, backgrounds, or real-life connections.

Decision Tree

A decision tree is a decision support tool that uses tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm.

2. LITERATURE SURVEY

In the era of big data, it is a phenomenon often appears that useful information is being submerged in a large number of useless information [12]. The data quality of Big Data has two problems: how to manage large-scale data and how to wash it. During the cleaning process, if the cleaning granularity is too small, it is easy to filter out the useful information; if the cleaning granularity is too common, it can't achieve the real cleaning effect. So between the quantity and quality, it requires careful consideration and weighed which is more evident in the real-time big data system. On the one hand, it requires the system to synchronize data in a very short time; on the other hand, it also requires the system to make a quick response to data in real time. The performance requirements of the speed of data transmission and data analysis are increasing. Moreover, the data may be filtered at a time node may become critical post processing data. Therefore, how to grasp the correlation between data and accurately determine the usefulness and effectiveness of data becomes a serious challenge.

Leimeister et al. [14] argue that the actors in the Cloud form a business value network moderately than a conventional business significance series. We identify the following actors in a Cloud-centric business value network (Figure 1): IT Vendors develop infrastructure software and operate infrastructure services; Service Providers develop and operate services; Service Aggregators offer new services by combining preexisting services; Service Platform Providers offer an environment for developing Cloud applications; Consulting supports customers with selecting and implementing Cloud services; Customers are the end-users of Cloud services.

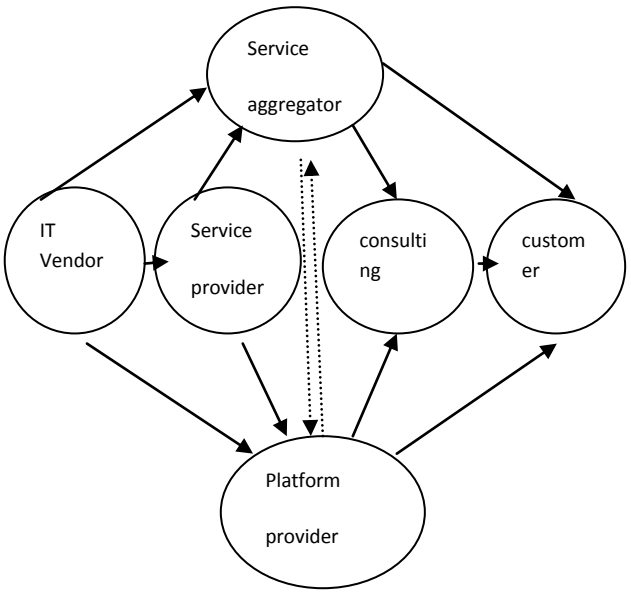


Figure-2: Cloud Actors and their Value Network

As that exploits the expression Infrastructure Provider for what they call IT Vendor. We deviate from to stress the fact that vendors that offer software that enables Cloud services should also be considered part of this actor group. They also use the expression Customer where others might utilize the term, Consumer. We decided to adhere to [14] in this case because service aggregators and service platform providers are consumers just as customers.

Big-Bench [13] is the modern attempt in the direction of planning big data benchmarks. BigBench focuses on big data offline analytics, thus accepting TPC-DS as the origin and adding up a top novel data types like semi un-structured data, as well as non-relational workloads. Even though BigBench has an entire exposure of data types, its object under test is DBMS and Map Reduce methods that declare to give big data explanations, guiding to limited exposure of software stacks. In addition at this time, it is not open-source for simple procedure and acceptance.

Chen et al. [15] found that application outlines from larger-scale Map Reduce Clusters organized in Facebook and Cloudera did not fit well-know statistical distributions. In this case, only real information can return the real system performances and workload features and hence the real world data is desired in big data benchmarks. On the other hand, to acquire real world information is a huge challenge because of two main explanations these are as follows: (1) the owner of real world data would not like to distribute their big data for dealing privacy and user privacies; and (2) although the real world data is offered on the internet it is undesirable for researchers to download terabyte-level data under the circumstance of existing internet traffic. Consequently, they present real-world data with two workloads under the authorization of our associate. These two workloads are stood and hot region, both of which are distinctive programs in our internal project associated to route data processing in the real world.

Aashish et al. analyze redundancy in the SPEC CPU2006 benchmark suite using micro-architecture metrics. They illustrate suggestion on a comparison of the benchmarks and reach your destination at significant subsets, and these subsets are representative of an extensive variety of applications areas without having many benchmarks with comparable features. The research consequence could clearly reduce execution time for system architecture researches [16]. Consequently, they concern the similar techniques to accomplish correspondence analysis in characteristic workloads in research domains in that order by micro-architecture level metrics. Alternatively, for a given scheme each workload in SPEC CPU2006 was performed as single process in a particular physical machine. Meanwhile, in our researchs, each MapReduce based workload was administration as a multiple process program in a spread computing atmosphere which consists of nine physical machines.

In this paper [1], they present a complete conversation of the BigBench measurement together with the database and the workload. In the development of extending BigBench they have acquired view from leading industry skilled about the significance in addition to entirety of the workload. After a methodological conversation of the benchmark and a conversation of illustration runs on two different "small" and "large" stages, they give a précis of the response in addition to designs for expectations expansions to the benchmark. They distinguish that Big Data is a difficult in addition to developing space. Big Bench characterizes only the initial step towards as long as a systematic method of benchmarking big data schemes. They anticipate that big data benchmarking will require being an agile movement for the near-term expectations, with the purpose of both maintain pace with changing scientific movements and the developing application conditions in this area.

3. PROPOSED METHODOLOGY

Here the proposed methodology is based on the combinatorial method of rules generation (fuzzy-Apriori) and classification (CART). The proposed methodology works in the following phases:

1 Training & Testing data
2 Apply fuzzy-Apriori algorithm on Input data
3 Then apply CART Algorithm to classification

Flow chart

Here flow chart of proposed methodology is based on the hybrid fuzzy-Apriori and CART classifier. Fuzzy-Apriori applies on the data and generates the frequent item sets and rules on the basis of support and confidence. CART classifier applies on frequent item sets and generates classification tree. Components of the flow chart are explained below:

Feature extraction based on attributes: Here we select a dataset from media database and calculate attributes by using features extraction method. Feature extraction method arranges the data sets in Attribute relation file format.

Apply Fuzzy-Apriori: Then we applying Apriori algorithm and generate rules from given dataset. After generating rules then we apply Apriori-fuzzy method on an over Apriori, because we are trying to generate minimum rules from given dataset before overall generating rules and frequent item set.

Apply cart classifier: After generating rules and frequent item set, then applying CART classifier technique to classify student opinion from generating item sets and finally generate a decision tree from given item set of the dataset.

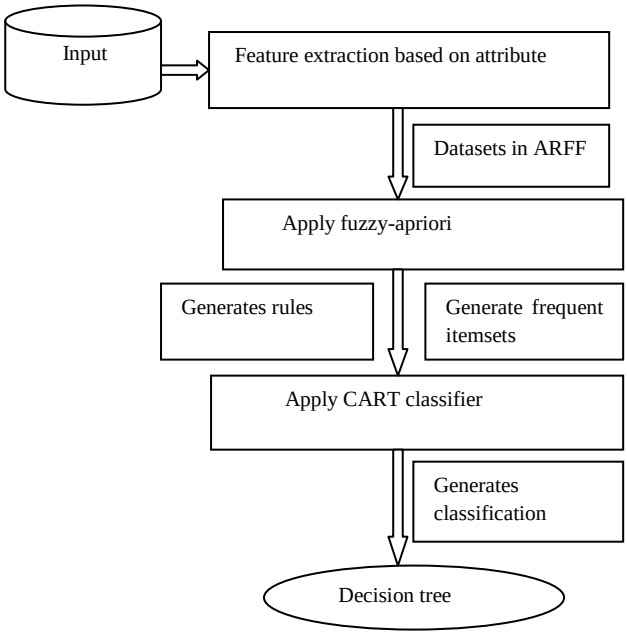


Figure 3. Flow Chart of the proposed Methodology

In this figure above proposed methodologies, here we select the dataset and calculate attributes by using features extraction method. Then we applying Apriori algorithm and generate rules from given dataset. After generating rules then we apply

Apriori-fuzzy method on an over Apriori, because we are trying to generate minimum rules from given dataset before overall generating rules and frequent item set. After generating rules and frequent item set, then applying CART classifier technique to classify student opinion from generating item sets and finally generate a decision tree from given item set of the dataset.

4. HYBRID (FUZZY-APRIORI & CART) ALGORITHM

The proposed algorithm is based on the combinatorial method of rules generation (fuzzy-Apriori) and classification (CART). Fuzzy-Apriori generates rules and frequent item sets. CART algorithm is applied on frequent item sets then generates the tree. The proposed algorithm works in the following phases:

1 Take an input dataset
2 Input Support & Confidence for the Apriori to generate Candidate Sets and rules.
3 For (k=1;$L_k \neq \emptyset$;k++) do begin
4 C_k=candidate generated from L_k;
5 For each transaction t in database do
6 Increment the count of all candidates in C_k that are contained in t
7 L_k=candidate in $C_k$with min_support
8 End
9 Return $U_k L_k$;
10 Traverse with each of the candidate sets and rules generated from Apriori.
11 For each itemset Lk in Apriori.
12 If it is frequent (based on Count[Lk]) over the whole dataset
13 Output (L_k)
14 Remove it
15 For each remaining itemsets L_k
16 Identify constituent singletons itemsets which are non-selected.
17 $s1, s2, \ldots \ldots sm$
18 Compute $X1 = \text{Min}(\text{Sup}(S1), \text{Sup}(S2))$
19 Compute $X2 = \text{Max}(\text{Sup}(S1), \text{Sup}(S2))$
20 $\text{Threshold} = \frac{X1+x2}{\text{total Transactions}} * \frac{1}{100}$
21 Compare threshold with the fuzzy activation thresholds
22 If threshold is greater
23 Itemset is Selected
24 Else
25 Itemset is Non-Selected.
26 Exit
27 Start at the root node of the candidate item sets
28 Compute Information of all the classes available in the dataset

$$I=\frac{-n(T)}{n(T)+n(F)} log\left(\frac{n(T)}{n(T)+n(F)}\right) - \frac{n(F)}{n(T)+n(F)} log\left(\frac{n(F)}{n(T)+n(F)}\right)$$

29 Compute Entropy for each attribute

$$E(A)=\frac{n(T)}{n(T)+n(F)} I(T,F) + \frac{n(F)}{n(T)+n(F)} I(T,F)$$

30 Compute Information Gain of each attribute
G=I-E
31 Update transaction tree with the node that is the highest Gain.

Example to Cover Algorithm

Take an example to implement the Algorithm

T-ID	A	B	C	D	E	F	G	H	I
T1	1	1	0	0	1	0	0	0	0
T2	1	1	1	1	1	0	1	1	0
T3	1	1	1	1	1	1	0	0	0
T4	1	1	0	0	1	0	0	0	0
T5	0	1	0	0	1	0	0	1	1
T6	0	1	0	1	1	1	1	0	0
T7	1	1	1	1	1	0	1	0	0
T8	1	1	1	1	1	1	0	1	0
T9	1	1	1	0	1	0	1	0	0
T10	0	1	1	1	1	1	0	0	1

Minimum Support = 30%

Here A=Author E=File

B=Age F=Data

C=Gender G=Friends

D=Trust Level H=No. of user connected

L1 Candidate Item Set

Item Set X	Supp(X)	Status
A	7/10=70%	Select
B	10/10=100%	Select
C	6/10=60%	Select
D	6/10=60%	Select
E	10/10=100%	Select
F	4/10=40%	Select
G	4/10=40%	Select
H	3/10=30%	Select
I	2/10=20%	Not Select

All Selected for the Next Candidate Item set since all have support greater than or equal to 30% accept I because its support value is less than 30%

L2 Candidate Item Set

Item Set X	Supp(X)	Status
AB	70%	Select
AC	50%	Select
AD	40%	Select
AE	70%	Select
AF	20%	Not Select
AG	30%	Select
AH	20%	Not Select
BC	60%	Select
BD	60%	Select
BE	100%	Select
BF	40%	Select
BG	40%	Select
BH	30%	Select
CD	50%	Select
CE	60%	Select
CF	30%	Select
CG	30%	Select
CH	20%	Not Select
DE	60%	Select
DF	40%	Select
DG	30%	Select
DH	20%	Not Select
EF	40%	Select
EG	40%	Select
EH	30%	Select
FG	10%	Not Select
FH	10%	Not Select
GH	10%	Not Select

This procedure follow up to L6 where Selected item called frequent and Non Selected item called Infrequent item set. Now we will apply Fuzzy-Apriori to make infrequent set to frequent.

Fuzzy Activation Threshold=0.01%

$$Threshold(\%)=\frac{min+max}{total\ transaction}*\frac{1}{100}$$

For L2 candidate set

Transaction	Threshold (%)	Status
AF	.11	Select
AH	.1	Select
CH	.09	Not Select
DH	.09	Not Select
FG	.08	Not Select
FH	.07	Not Select
GH	.07	Not Select

We have to follow this procedure up to L6 data set and check for the Frequent data set.

NO. of repetition time of selected Frequent Item Set

A	B	C	D	E	F	G	H
71	65	55	66	72	45	50	49

Total set of rule selected=139

INFORMATION-

$$I(T/F) = \frac{-n(T)}{n(T)+n(F)} log\left(\frac{n(T)}{n(T)+n(F)}\right) - \frac{n(F)}{n(T)+n(F)} log\left(\frac{n(F)}{n(T)+n(F)}\right)$$

T=79, F=60, I(T/F)=0.296

ENTROPY-

$$E(A) = \frac{n(T)}{n(T)+n(F)} I(T,F) \frac{n(F)}{n(T)+n(F)} I(T,F)$$

E(A) = 0.200 (T=40, F=31)

E(B) = 0.286 (T=33, F=32)

E(C) = 0.290 (T=30, F=25)

E(D) = 0.294 (T=35, F=31)

E(E) = 0.293 (T=37, F=35)

E(F) = 0.170 (T=24, F=21)

E(G) = 0.262 (T=35, F=15)

E(H) = 0.291 (T=29, F=20)

GAIN-

G(A) = I(A)-E(A) =0.096

G(B) = 0.01

G(C) = 0.006

G(D) = 0.002

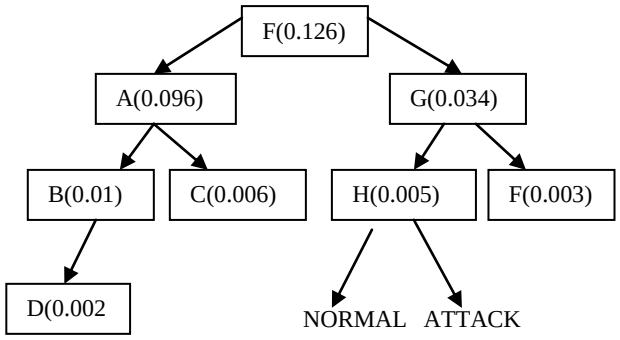
G(E) = 0.003

G(F) = 0.126

G(G) = 0.034

G(H) = 0.005

Here highest gain of G(F) =0.096 will be at the root and remaining are on leaf



If a message comes with the following value

A	B	C	D	E	F	G	H	
0.2	0.38	0.25	0.8	0.56	0.19	0.023	0.056	

Check F=0.19 compare with decision tree choose right branch.

1. Now at right branch G is the root.
2. Check G=0.0023 compare with the decision tree choose left branch.
3. Now at left branch H is the root.
4. Compare H=0.056 compare with the decision tree choose right branch.
5. Now at the right branch Decision is taken.
6. Hence type of this message will be attack.

Hence we easily conclude from the decision tree that the message type will normal or abnormal. Here from this example we see this message type will be attack or abnormal that's why this message will not considered as a common behaviour of students.

5. RESULT ANALYSIS

Table 1. Analysis of Accuracy

	Accuracy	
Probability Threshold	**Existing Work**	**Proposed Work**
0.1	0.6223	0.6837
0.2	0.663	0.712
0.3	0.6879	0.7328
0.4	0.6996	0.752
0.5	0.7019	0.7628
0.6	0.7052	0.793
0.7	0.7064	0.81
0.8	0.706	0.818
0.9	0.7078	0.823
1	0.7088	0.834

Table 2. Analysis of Precision

	Precision	
Probability Threshold	**Existing Work**	**Proposed Work**
0.1	0.6266	0.6523
0.2	0.6675	0.6819
0.3	0.6934	0.721
0.4	0.7068	0.7384
0.5	0.7091	0.7509
0.6	0.714	0.7712
0.7	0.7152	0.7833
0.8	0.7158	0.7945
0.9	0.7176	0.82
1	0.7199	0.827

Table 3. Analysis of F-Measure

	F-Measure	
Probability Threshold	**Existing Work**	**Proposed Work**
0.1	0.7099	0.7196
0.2	0.7214	0.7244
0.3	0.7262	0.7370
0.4	0.7260	0.7309
0.5	0.7189	0.7347
0.6	0.7172	0.7358
0.7	0.7153	0.7357
0.8	0.7135	0.7340
0.9	0.7138	0.7386
1	0.7143	0.7347

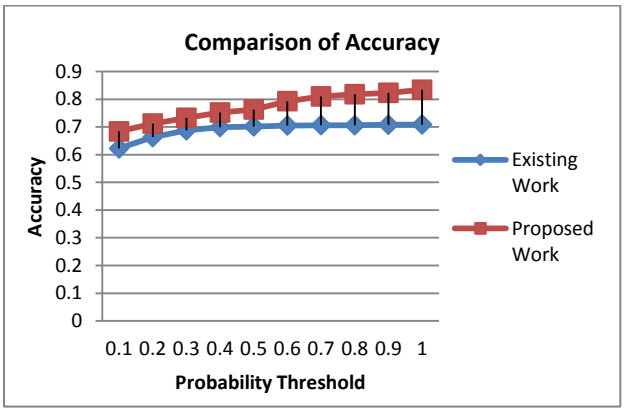


Figure 1. Comparison of Accuracy

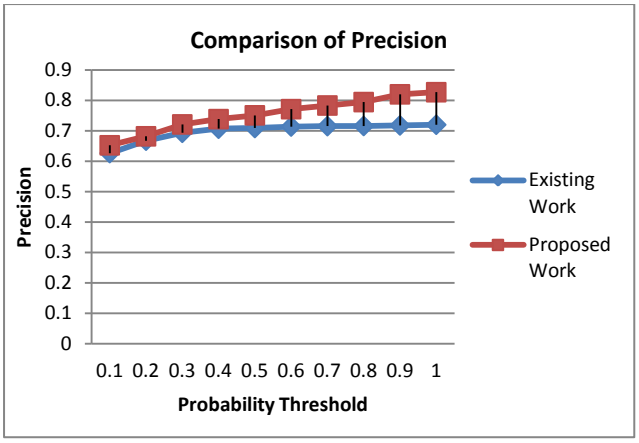


Figure 2. Comparison of Precision

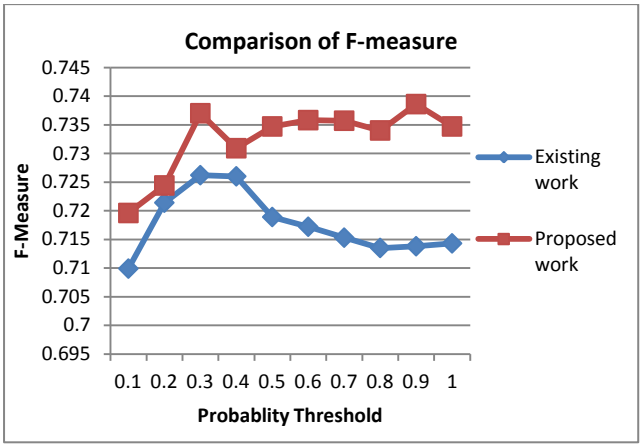


Figure 3. Comparison of F-Measure

6. CONCLUSION

Social network analysis (SNSs) is the learning of human inter-connectivity from social media websites are those instruments which have been produced, to facilitates easier and more productive communication within an insecure channel in these open network. The meadow of social media network investigation has come into sight over the past two years. Social network analysis (SNSs) is the learning of human inter-connectivity from social media websites are those instruments which have been produced, to facilitates easier and more productive communication within an insecure channel in these open network. The meadow of social media network investigation has come into sight over the past two years as a possible means of organizing the opinion of individual and groups of person, as the concern to scrupulous proceedings.

The Proposed Methodology implemented here for the Classification of Student's Learning Experience on Social Media Datasets such as Twitter and Facebook. The Methodology implemented provides better accuracy and Precision as compared to the existing Naïve Bayes Multi-label classifiers. The Methodology implemented can also better explains various Emotions and Sentiments of Students about their Educational Learning Experience.

7. REFERENCES

[1] Z. Fadika, E. Dede, M. Govindaraju, and L. Ramakrishnan. Benchmarking MapReduce implementations for application usage scenarios. In GRID 2011

[2] Yadagiri S, Thalluri P V S. Information technology on surge: information literacy on demand. DESIDOC Journal of Library & Information Technology, 2011, 32(1):64-69.

[3] Cohen J, Dolan B, Dunlap M, Hellerstein JM, Welton C. MAD skills: New analysis practices for big data [J]. PVLDB, 2009, 2(2):14811492.

[4] Randal E. Bryant & Joan Disney. Data-Intensive Supercomputing: The case for DISC [R].2007.10: 1-14.

[5] John Boyle. Biology must develop its own big-data systems.Nature. 2008, 499(7): 7.

[6] Wang Yuan-Zhuo, Jin Xiao-Long, Chen Xue-Qi. Network Big Data: Present and Future [J].Chinese Journal of Computer. 2013, 36(6):1125-1138.

[7] B. Cooper et al. Benchmarking cloud serving systems with yes. In SOCC 2010.

[8] Z. Fadika, E. Dede, M. Govindaraju, and L. Ramakrishnan. Benchmarking MapReduce implementations for application usage scenarios. In GRID 2011

[9] M. Ferdman et al. clearing the clouds, a study of emerging scale-out workloads on modern hardware. In ASPLOS 2012.

[10] Lei Wang, Jianfeng Zhan, ChunjieLuo, "BigDataBench: a Big Data Benchmark Suite from Internet Services" High-Performance Computer Architecture (HPCA), IEEE 20th International Symposium on2014.

[11] Yadagiri S, Thalluri P V S. Information technology on surge: information literacy on demand. DESIDOC Journal of Library & Information Technology, 2011, 32(1):64-69.

[12] Cohen J, Dolan B, Dunlap M, Hellerstein JM, Welton C. MAD skills: New analysis practices for big data [J]. PVLDB, 2009, 2(2):14811492.

[13] Randal E. Bryant & Joan Disney. Data-Intensive Supercomputing: The case for DISC [R].2007.10: 1-14.

[14] John Boyle. Biology must develop its own big-data systems.Nature. 2008, 499(7): 7.

[15] Wang Yuan-Zhuo, Jin Xiao-Long, Chen Xue-Qi. Network Big Data: Present and Future [J].Chinese Journal of Computer. 2013, 36(6):1125-1138.

[16] B. Cooper et al. Benchmarking cloud serving systems with yes. In SOCC 2010.