Open access Journal International Journal of Emerging Trends in Science and Technology
Impact Factor: 2.838 DOI: http://dx.doi.org/10.18535/ijetst/v3i09.12

Academic Recommendation

Authors
Shikha Agarwal[1], C.P.Singh[2]
[1]M.Tech Computer Science &Engineering, SRGI, Jhansi
[2]Asst. Prof. Computer Science & Engineering, SRGI, Jhansi

Abstract
The proposed work is intended to develop a semi-automatic technique for classifying the sentiments based text. Basically in this presented work for text classification the decision trees are implemented which are the supervised learning algorithms. But the additional effort for tagging of data is necessary therefore that is known as the semi supervised model. The technique is applied on the online student's learning and discussion to find the students experience and improve the experience in further learning processes. Therefore the entire system development is performed on two major modules first for generating the student's communication data and then uses it with the supervised model for text classification according to the user emotions. In this technique the web data is first pre-processed for improving the quality of learning data. After that the data is used with the NLP parser for finding the communicated text features. The computed text features are than used with two different supervised learning models namely C4.5 and the ID3. These models are basically a kind of decision trees, during training of these algorithms the algorithm generates the tree. These generated decision are termed here as the trained model. The trained classifier is further used for real time classification of communicated text for the binary classification. The implementation of the entire text mining concept is performed using the JAVA technology and the classifiers performance is compared using the accuracy, error rate, memory consumption and time consumption. According to the computed classifiers performance the proposed technique namely c4.5 based classifier perform more accurate classification as compared to the traditional ID3.

Keywords— *text classification, emotional classification, NLP, machine learning, implementation, binary classification.*

1. Introduction

The data mining is a rich domain of information mining and extraction of knowledge. That is offering the tools and techniques that are works automatically to analyse the significant amount of data. Therefore these techniques are utilizes the different computational algorithms which estimates the patterns of data during the algorithms training. According to the algorithms training the data models are divided in two main categories supervised and unsupervised learning techniques. In literature it is found that the supervised learning techniques are much accurate and efficient as compared to the unsupervised learning techniques. Therefore in this presented work the supervised learning approaches are targeted to study for classifying the text data according to the sentiments and their orientations.

In this proposed work for students emotions are and their learning experience is tried to recover. Therefore in first a student's learning and assignment's portal is prepared and then the generated data by students and the faculty members are analysed for the binary classification scheme for finding the student's emotions or sentiments. Therefore there are two different modules of the application is designed first a web application which is used to generate the data of students and their assignments with the user's reactions. In next step a desktop application is designed for extracting the web application data. the desktop application is also responsible for pre-process the data and perform training and testing of the extracted text communications between different users.

The proposed work is intended to enhance the previously existing model of social media based student learning experience using the opinion mining technique. The main aim of the work is to design effective and accurate orientation mining technique for the social network user. In order to work with the opinion mining domain the following issues related to the previous work is recovered for the improvement some of the issues and challenges to optimize the existing system is given as.

1) Not all students are associated or active in discussing issues and solutions in social networking.
2) The entire system modelling is performed on an engineering students group, the involved not all students are discussing good things, thus for finding the issues, solutions and other aspects of data. Too fewer amounts of data is available for analysis good things and the study experience.
3) Author able to recover only small amount of significant amount of data according to the given theme of data analysis. There are also some additional information is hidden in the data which is not recovered.
4) There are not a mathematical model exist that helps to identify the real world problem with the student behaviour. Thus an identical theme is required on which the student’s issues are required to correlate not for all the real world issues.

In order to recover most of issues from the previously obtained issues the following objectives are included to design a newer solution.

1) Prepare the social networking for the students and faculties, on which the students and teachers can share their data and experience.
2) That is a place where all the students are actively participating on accessing lecture notes, discuss their problems and find the solutions.
3) Oriented to the specific theme on which the issues of students and the solutions of the issues can be solvable.
4) The proposed model includes the text classification models for improving the classification performance of social networking based text analysis.
5) Only the correlation among the students study relevant issues are correlated to the mathematical model by which the proposed model becomes more effective.

2. Proposed Work

The section provides the entire details about the approach and the solution design technique. Therefore first include the detailed overview of the domain on which the proposed work is targeted, then after the used algorithms are described and finally the methodology of solution development is also reported.

A. Domain overview

The data mining is a domain for the extraction of knowledge from the raw set of data. Therefore the mining techniques are worked on the basis of their learning techniques there are two kinds of learning techniques are available supervised and unsupervised learning. The unsupervised learning techniques support the classification of data and the unsupervised learning technique supports the cluster analysis of data. But for the pattern detection and classification the supervised learning algorithms are used. In supervised learning the computational algorithms are works on the training samples first to estimate the patterns and then these trained models are used for classification and pattern recognition. But to train the data models needs the set of attributes and their class labels. These class labels are help to recognize the similar pattern attributes.

In this presented work the decision tree classifier is used for the learning the communications of the students. Thus first the students communications are recorded and then these records are parsed using the NLP tool. The NLP tool helps to convert the data into the defined part of speech. Using the part of speech the set of attributes are prepared and used with the decision tree algorithm to develop the classification data model. There are two different decision trees namely ID3 and C4.5 algorithm is applied on the training dataset. This data model is used to classify the user communication in web portal for finding the students good or bad experience. This section provides the overview of the proposed concept of the proposed technique of sentiment based text analysis next section provides the algorithm details.

B. Algorithm study

This section provides the detailed understanding of the different algorithms which are used for implementation and system design.

Classical C4.5 Algorithm

C4.5 (developed by Quinlan, 1993) an algorithm that learns the decision-tree classifiers, It has been observed that C4.5 performs short in the domain where there is pre-entrance of continuous attributes compared with the learning tasks with mostly separate attributes. For instance, a system which looks for well-defined decision tree with 2 levels and then put comments [34]:

"The accuracy of trees made with T2 is equalized or even exceed trees of C4.5 upon 8 out of all the datasets, with the entire except one that have incessant attributes only."

INPUT: An exploratory data set of data (D) portrayed with the means of discrete variables.

OUTPUT: A decision tree say T which is constructed by means of passing investigational data sets.

1) A node (X) is created;
2) Check if the instance falls in the same class.
3) Make node (X) as the leaf node and assign a label CLASS C;
4) Check IF the attribute list is empty, THEN
5) Make node(X) a leaf node and assign a label of most customary CLASS;
6) Now choose an attribute which has highest information gain from the provided attribute List, and then marked as the test_attribute;
7) Confirming X in the role of the test_attribute;
8) In order to have a recognized value for every test_attribute for dividing the samples;
9) Generating a fresh twig of tree that is suitable for test_attribute = att$_i$from node X;
10) Take an assumption that Bi is a group of test_attribute=att$_i$in the samples;
11) Check If Bi is NULL, THEN
12) Next, add a new leaf node, with label of the most general class;
13) ELSE a leaf node is going to be added and returned by the Generate_decision_tree.

ID3 Decision Tree Basics

Engineered by Ross Quinlan the ID3 is a straightforward decision tree learning algorithm. The main concept of this algorithm is construction of the decision tree through implementing a top-down, greedy search by the provided sets for testing every attribute at each node of decision. With the aim of selecting the attribute which is most useful to classify a provided set of data, a metric is introduced named as Information Gain [35].

To acquire the finest way for classification of learning set, one requires to act for minimizing the fired question (i.e. to minimize depth of the tree). Hence, some functions are needed that is capable of determine which questions will offer the generally unbiased splitting. One such function is information gain metric.

ENTROPY

In order to define information gain exactly, we require discussing entropy first. Let's assume, without loss of simplification, that the resultant decision tree classifies instances into two categories, we'll call them P (positive) and N (negative)

Given a set E, containing these positive and negative targets, the entropy of S related to this Boolean classification is:

Entropy(E)= –P(positive) $\log_2$P (Positive) – P (negative) $\log_2$P(negative)

P (positive): proportion of positive examples in set E

P (negative): proportion of negative examples in set E

INFORMATION GAIN

As already discussed, for cutting down the depth of a decision tree, while traversing the same, selection of the best possible characteristic is mandatory in order to split the tree, this clearly shows that attribute with minimum drop of entropy will be the superlative pick.

Here, the information gain can be termed as required drop in entropy in relation with individual attribute during the decision tree splitting.

The information gain, Gain (E, A) of an attribute A,

Gain(E,A)= Entropy(s) $\sum_{i=1}^{v} \frac{Ev}{E} \times Entropy(Ev)$

This concept of gain can be utilized to decide positions of attributes as well as to construct decision trees in which every node is positioned the attribute with maximum gain amongst those attributes that are not considered in the path from the root yet.

The intention of this ordering is:

1) To generate small sized decision trees in order to identify records after only a handful decision tree splitting steps.
2) To attain the desired level of unfussiness of the decisional approaches.

The main points of ID3 algorithm are as follows:

1) Obtain all idle attributes and calculate their entropy relating to test samples
2) Prefer that attribute Which has least entropy (or, consistently, the highest Information gain)
3) Create a node having such attributes.

The algorithm is as follows:

Table 2.1 ID3 Decision Tree

Input: Examples, Target_Attribute, Attributes
Output: Decision Tree
Process:

```
Process:
  ▪ Produce a node being root node of the tree
  ▪ Check if all the examples are positive, If yes
    then generate a single node tree, ROOT, having
    label = +.
  ▪ In case all the examples are found negative,
    then make a single node tree, ROOT, with label
    = -.
  ▪ If there are no attributes for prediction, then
    create a single node tree ROOT labelled as the
    most ordinary used value for that attribute.
  ▪ Else Start following procedure
      ▪ M = an attribute that is classifying the
        Examples in Best way.
      ▪ Make M the decision tree attribute
      ▪ Repeat for every probable value, $v_i$, of M,
          ▪ Expand the Root with one branch,
            equivalent to the test M = $v_i$.
          ▪ Let Examples($v_i$) be a subset of
            examples that have the value $v_i$ for
            M
          ▪ If Examples($v_i$) is empty
              • Add a leaf node under
                the new branch, labelled
                with most ordinary value
                of the attribute in the
                examples
          ▪ Otherwise add the sub-tree ID3
            (Examples($v_i$), Target_Attribute,
            Attributes – {A}), under this new
            branch
  ▪ End
  ▪ Return Root
```

C. Methodology

The figure 2.1 and figure 2.2 shows the working of the proposed solution for extracting information from the web portal and utilizing them to analyse the student's experience. The used components of the given system are also discussed in this section.

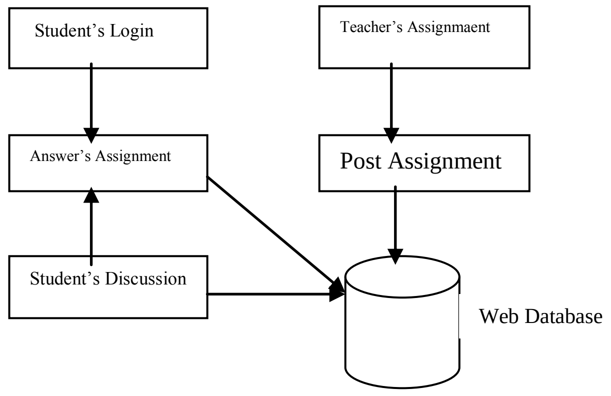

Figure 2.1 web application management

In figure 2.1 the key points of the proposed web application is highlighted. Therefore the role of students and teachers are demonstrated for the data generation phase. In this scenario first the ***teacher is login*** to the system; it is a basic authentication for identifying the teacher and managing their session. After login the teacher can ***post an assignment*** for the targeted students. On the other hand as the ***student login*** to the system he/she find the list of assignments on their screen. Students ***write the answers*** for the assignment questions online. In addition of that the student can also participate in ***discussion***. All the communication is performed using the data base scenarios therefore each entry is preserved on the database. That data base is termed as the web database. This data base is further used with the prepared classification model. The web data base is further used with the figure 2.2 in this diagram the first the ***pre-processing*** technique is called. Using this technique the data is refined and the special characters and stop words are removed first. This process reduces the amount of data and makes it clean for further use with the algorithm.

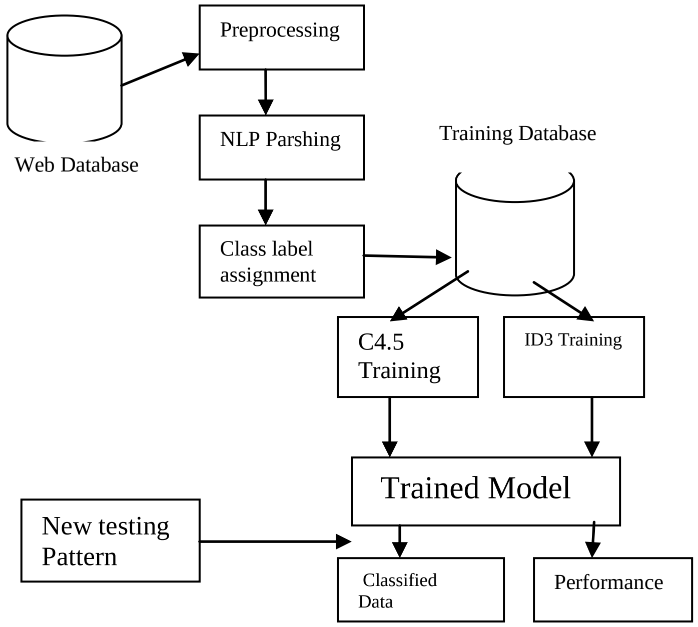

Figure2.2 Sentiment Based Classification

The pre-processed data is used with the ***NLP (natural language processing) parser***. That parser parses the entire data into the part of speech fractions. These fractions are now computed and aggregated as the attributes additionally a human effort is need to provide the class labels in the phase of the ***class label assignment***. After assigning the class labels on the selected training samples the data is stored on a temporary database termed here as the ***training database***. The training data base contains a set of instances with the attributes and their classes, these patterns are first used with the classifiers implemented namely ***ID3 and C4.5***. These classifiers are performing the training over the data and prepare the decision tree model. The developed tree using the training database data is termed here as the ***trained model***. The trained model is now able accept the new or upcoming attributes as input for ***classifying them*** according to their class labels. at the same time the performance of the models are also computed in terms of accuracy, error rate, memory and time consumption. This section defines the working of the proposed model and the next section provides the summarized step of the involved processes.

D. Proposed algorithm

This section provides the algorithm steps of the proposed data model using C4.5 algorithm for classifying the data patterns into their relevant sentiments and their patterns.

Table 2.2 proposed algorithm

```
Input: unlabelled data D, user input training classes $C_u$, Test Set input T
Output: classes C

Process:
    1. $R_d$= readData(D)
    2. P  = preProcessData($R_d$)
    3. For each line in Datatset do
           a. Ai = NLP.Parse(Li)
           b. Ii  = CreatePattern(Ai,Cu)
    4. End for
    5. $D_{b=}$ DB.Store ($I_i$)
    6. Tmodel=C4.5· train(Db)
    7. C=$T_{model}$ ·test(T)
    8. Return C
```

3. Results Analysis

The given section provides the study about the evaluation of the proposed classification algorithm and the comparative performance study with the traditional classification algorithm. The obtained performance factors and their approximated values are given.

A. Accuracy

The accuracy is a measurement of the data model for finding the amount of correctly classified data using the input samples. The performance of the algorithm in terms of accuracy can be evaluated using the following formula.

$$\mathbf{accuracy\%} = \frac{\textit{total correctly classified data}}{\textit{total input datasets}} \times \mathbf{100}$$

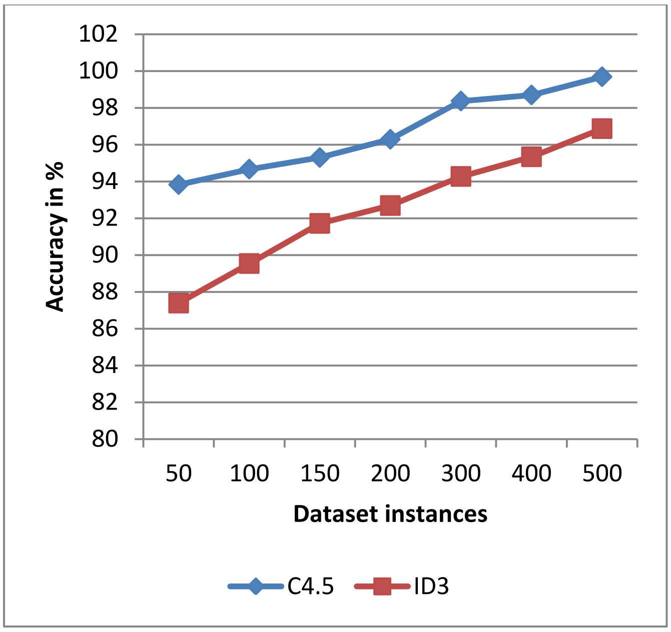

Figure 3.1 accuracy

The performance of the proposed c4.5 classifier and the traditional ID3 algorithm is compared using the figure 3.1 and the table 3.1.

Table 3.1 accuracy

Dataset size	C 4.5	ID3
50	93.82	87.38
100	94.66	89.53
150	95.29	91.71
200	96.28	92.68
300	98.36	94.26
400	98.69	95.33
500	99.68	96.86

In this diagram the X axis shows the training samples in the dataset and the Y axis shows the obtained accuracy in terms of percentage. The results of both the classifiers are demonstrating the different behaviour of classification aspects, in the traditionally implemented classifier the performance of the classification is reduces as the amount of training instances are increases. On the other hand the performance of the proposed classification technique is increases as the amount of training samples are increase. Thus the proposed classifier performs more effectively as compared to traditional manner of classification. For analysing the results in the statistical manner the mean accuracy of both the classifiers are computed and their difference in performance is reported using the figure 3.2. In this diagram the mean performance of both the method in terms of accuracy is demonstrated using the Y axis and the X axis contains the implemented methods for making comparative performance study. According to the obtained performance the proposed classifier is producing approximately 96% of accurate results and the traditional classifier produces the 92.53 % of accurate results. Thus the proposed classification technique is much efficient and accurate as compared to the traditional technique of data classification.

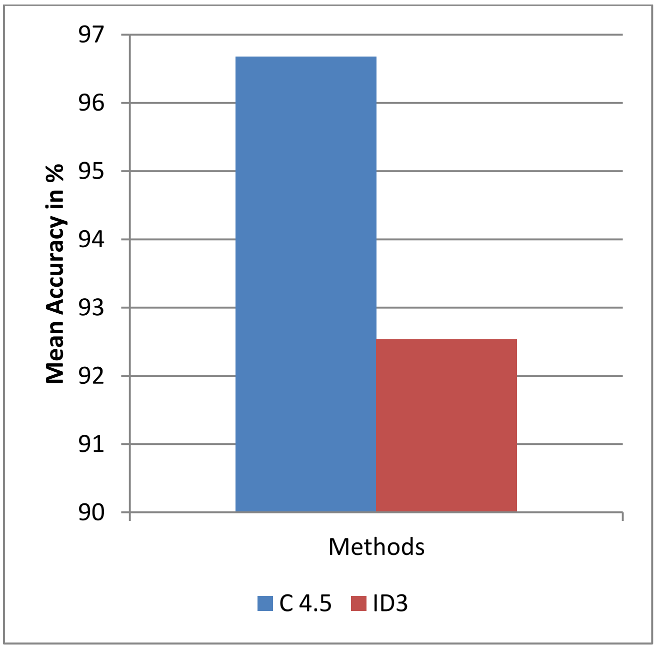

Figure 3.2 mean accuracy

B. Error rate

The error rate of the classifier provides the estimation about the misclassified samples during the testing of the trained classifier. The evaluation of error rate can be performed using the following formula.

$$\text{error rate\%} = \frac{total\ misclassified\ samples}{total\ input\ samples} \times 100$$

Or

error rate %= 100 – accuracy%

The comparative error rate of the proposed and traditional classification technique is provided using the table 3.2 and the figure 3.3. The given figure includes the X axis to show the size of training samples and the Y axis shows the amount of misclassified patterns in terms of percentage. According to the demonstrated results the error rate of the proposed classifier is reduces as the amount of training instances are increases in the database. On the other hand the error rate of the traditional scheme is increases as the amount of data for learning is increases. Thus the proposed classifier is improving the outcomes of the classification with increasing the learning patterns.

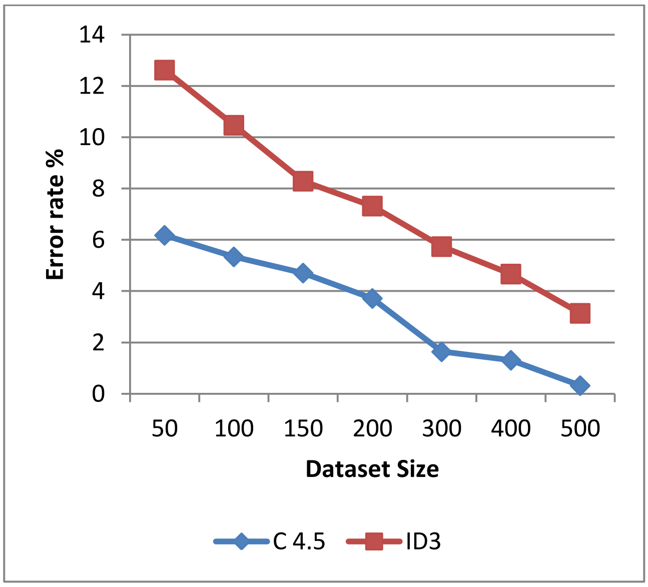

Figure 3.3 error rate

Table 3.2 error rate

Dataset size	C 4.5	ID3
50	6.18	12.62
100	5.34	10.47
150	4.71	8.29
200	3.72	7.32
300	1.64	5.74
400	1.31	4.67
500	0.32	3.14

In order to understand the performance of the classification more clearly the mean error rate percentage is evaluated and reported using the figure 3.4. In this figure the amount of error rate produced by the algorithms are demonstrated using Y axis and X axis shows the methods implemented with the system.

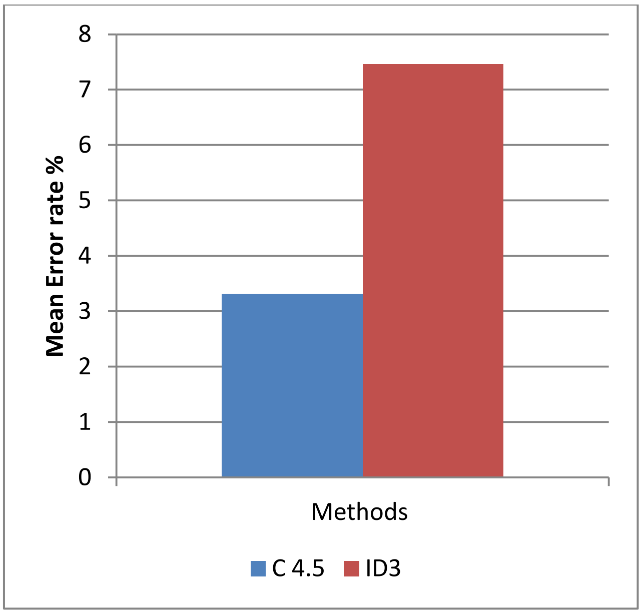

Figure 3.4 mean error rate

According to the obtained results the from the mean error rate percentage the proposed C4.5 produces more effective and improving performance as compared to the traditional ID3 classification technique.

C. Memory usages

The amount of main memory required to successfully execute the algorithms is known as the memory consumption of the algorithms. The given figure 3.5 and the table 3.3 show the comparative performance of both the implemented classifiers. In the given diagram the X axis shows the number of training input samples produced for the training to the data models and the Y axis shows the amount of main memory consumed by the implemented algorithms. According to the obtained results the amount of memory consumption in the proposed data modeling is higher as compared to traditional technique because the proposed classifier needs to process the data using both the classifiers.

Table 3.3 memory usage

Dataset size	C4.5	ID3
50	27817	25818
100	29488	26891
150	31938	27716
200	33462	28612
300	35938	29981
400	37362	30164
500	39882	31332

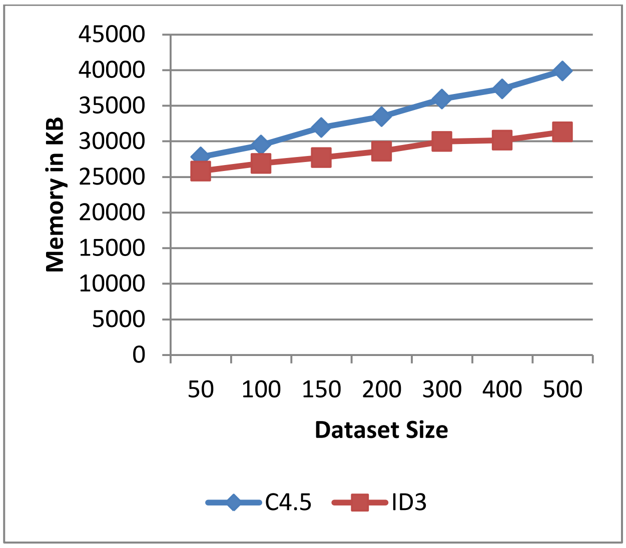

Figure 3.5 memory usage

In order to understand the memory usage difference among both the classification technique the mean memory consumption is demonstrated using the figure 3.6 in this diagram the X axis shows the amount of instances of the data used for training and the Y axis shows the amount of main memory consumed during evaluation of data.

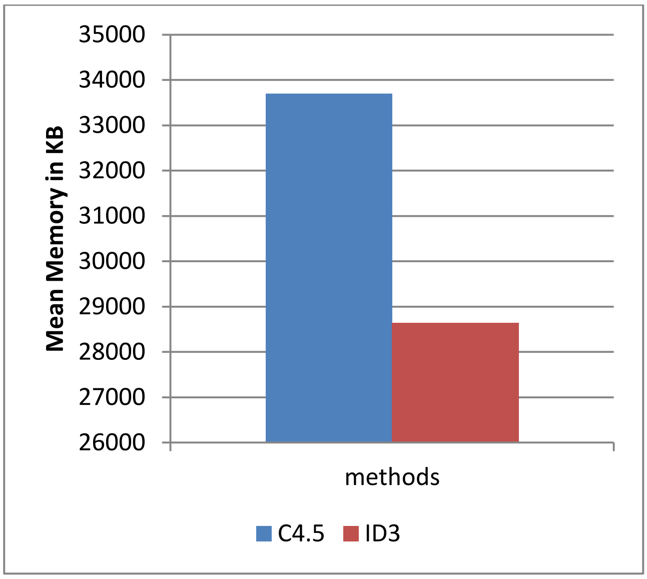

Figure 3.6 mean memory consumption

Time consumption

The amount of time required to process the data using the proposed algorithm is termed here as the time consumption of the system. The comparative time consumption of both the data models during the training is demonstrated using table 3.4 and figure 3.7. In this diagram the X axis contains the amount of data used for training and the Y axis shows the amount of time required to process the data samples. According to the obtained results the proposed technique consumes higher time as compared to the traditional classifier. The proposed scheme utilizes the back propagation neural network and learning of this algorithm is an iterative process thus the amount of time is higher as compared to ID3.

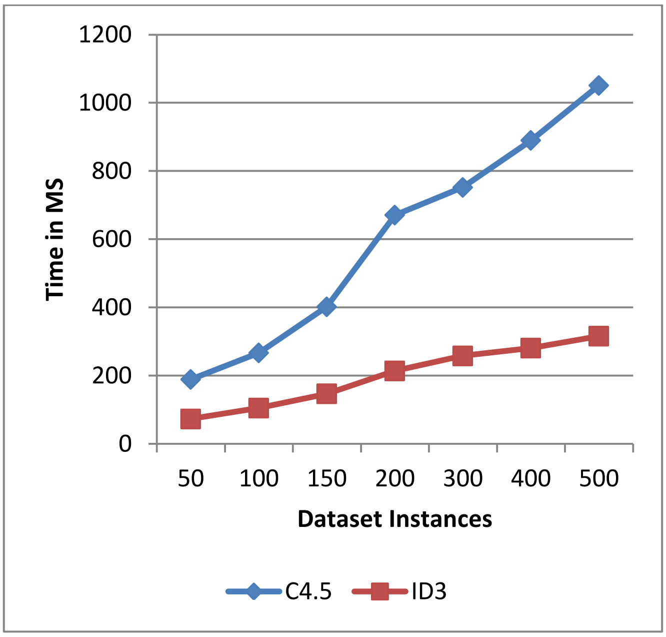

Figure 3.7 time consumption

Table 3.4 time consumption

Dataset size	C4.5	ID3
50	189	73
100	267	105
150	402	147
200	671	214
300	752	258
400	890	281
500	1051	316

In order to understand the difference among both the technique's performance the mean time consumption of both the algorithms are computed. According to the obtained performance the proposed technique consumes more time as compared to the traditional technique. Thus the proposed model is a time consuming model for training time.

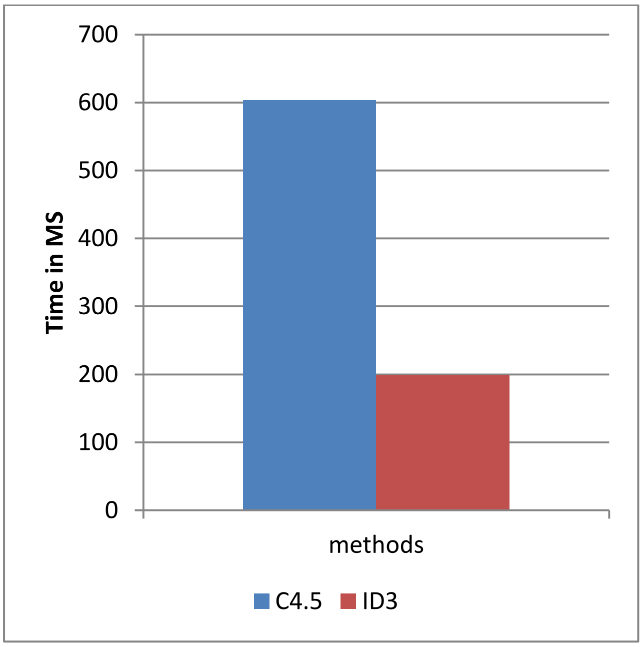

Figure 3.8 mean time consumption

4. Conclusions

The proposed work is intended to find the student orientation for the study load and their learning experience. Therefore a data mining based sentiment analysis technique is implemented on the assignment portal to classify the student's communications. This chapter provides the summary of the entire efforts made additionally the future extension of the work is also suggested.

A. Conclusion

The data mining and their technique offers to analyse the data and recover the patterns to use with the different kinds of applications such as banking, business, advertisement agencies and others. The data mining approaches are basically the computational algorithms. That is first takes the training on the predefined data and then use to recover the target patterns. According to the learning techniques the data models are defined as supervised and unsupervised approaches. The supervised techniques are more efficient and accurate for the pattern classification and recognition. Therefore the supervised approach is used for classifying the data in the proposed work.

In this presented work the classification technique is demonstrated by implementing with the sentiment based classification of text. Therefore the proposed work is motivated from the text classification techniques and their investigation. In order to do this task the student's assignment portal data is used for analysis and their classification. Thus first of all a web portal with three different user access namely student, faculty members and the administrator is implemented using java technology. After that the user's communication is captured and stored in the data base. After preparing the data communicated between different users a separate module of desktop application is created. That application utilizes the data in three modules. In first the pre-processing is taken place where the unwanted characters and the stop words are removed first. After that the NLP processing tool is used to find the part of speech of all the communicated sentences. The parsed sentences are label with the 0 and 1 classes where the 0 means the negative response of the student and 1 is for the positive response. After pre-processing of the data the refined data is used with the classifiers for performing the training.

In this phase there are two classifiers namely C4.5 and ID3 is implemented. The training on the pre-processed data is performed for both the implemented classifiers. After training the classifiers are prepared to classify the data. Now the new arrived patterns are produces as input to the trained data models and the classifiers are classifies the given patterns. The implementation of the desktop application is also performed using the JAVA technology and their performance in terms of accuracy, error rate, memory consumption and time consumption is computed. The performance summaries of the implemented techniques are given using table 4.1.

Table 4.1 performance summary

S. No.	Parameters	C4.5	ID3
1	Accuracy	High	Low
2	Error rate	Low	High
3	Memory	High	Low
4	Time consumption	High	Low

According to the obtained performance the proposed classifier for sentiments analysis perform accurately but the time and memory consumption of the classifier is higher as compared to the traditional approach of classifier. Therefore the proposed classification technique for sentiments analysis is referred for those applications where the accuracy is considerable as compared to resources.

B. Future extension

The main objective of the proposed work is to find the students orientation according to their interest and learning is accomplished successfully. In further the following improvements are suggested for future work.

1) Currently the proposed technique limited for classifying the text in two classes need to improve the classification techniques for more than two classes.
2) Resource consumption during accurate classification is higher as compared to traditional approach thus need to improve the classifier performance to improve the time and space complexity

References

1. Xin Chen, mihaelavorvoreanu, Krishna madhavan, “mining social media data for understanding students learning experience”, IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, VOL. 7, NO. 3, JULY-SEPTEMBER 2014
2. Data Mining: What is Data Mining?, http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
3. Data Mining - Applications & Trends, http://www.tutorialspoint.com/data_mining/dm_applications_trends.htm
4. Mahak Chowdhary, ShrutikaSuri and Mansi Bhutani, “Comparative Study of Intrusion Detection System”, 2014, IJCSE All Rights Reserved, Volume-2, Issue-4
5. Mrs. PradnyaMuley, Dr. Anniruddha Joshi, “Application of Data Mining Techniques for Customer Segmentation in Real Time Business Intelligence”, International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163, Issue 4, Volume 2 (April 2015)
6. Ritika, “Research on Data Mining Classification”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 4, April 2014
7. ABBAS JAFARI, S.S.PATIL, “USE OF DATA MINING TECHNIQUE TO DESIGN A DRIVER ASSISTANCE SYSTEM”, Proceedings of 7th IRF International Conference, 27th April-2014, Pune, India, ISBN: 978-93-84209-09-4
8. FabricioVoznika, LeonarDoviana, “Data Mining Classification”, http://courses.cs.-washington.edu/courses/csep521/07wi/prj/leonardo_fabricio.pdf
9. A.K. Jain, M.N. Murthy, P. J. Flynn, “Data Clustering: A Review”, © 2000 ACM 0360-0300/99/0900–0001
10. A Comparative Study of Data Clustering Techniques, Khaled Hammouda, Dept of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1
11. Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek, "Density-based Clustering", WIREs Data Mining and Knowledge Discovery 1 (3): 231–240. doi:10.1002/widm.30
12. B. V. Rama Krishna, B. Sushma, “Novel Approach to Museums Development & Emergence of Text Mining”, ISSN 2249-6343, International Journal of Computer Technology and Electronics Engineering (IJCTEE), Volume 2, Issue 2

13. H. P. Luhn, “A Business Intelligence System”, Volume 2, Number 4, Page 314 (1958), Nontopical Issue, IBM Research Journals
14. Andreas Hotho, Andreas Nurnberger, Gerhard Paaß, FraunhoferAiS, “A Brief Survey of Text Mining”, Knowledge Discovery Group Sankt Augustin, May 13, 2005s
15. Hien Nguyen, Eugene Santos, and Jacob Russell, “Evaluation of the Impact of User-Cognitive Styles on the Assessment of Text Summarization”, IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, Vol. 41, No. 6, November 2011
16. Umajancy. S, Dr. Antony Selvadoss Thanamani, “An Analysis on Text Mining –Text Retrieval and Text Extraction”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8, August 2013
17. MilošRadovanović, MirjanaIvanović, “Text Mining: Approaches And Applications”, Abstract Methods and Applications in Computer Science (no. 144017A), Novi Sad, Serbia, Vol. 38, No. 3, 2008, 227-234
18. S. L. Pandhripande and Aasheesh Dixit, “Prediction of 2 Scrip Listed in NSE using Artificial Neural Network”, International Journal of Computer Applications (IJCA), Volume 134, No.2, January 2016.
19. Dr. B. Srinivasan and K. Pavya, “A STUDY ON DATA MINING PREDICTION TECHNIQUES IN HEALTHCARE SECTOR”, International Research Journal of Engineering and Technology (IRJET), PP. 552-556, Volume 3, Mar-2016
20. Vipin Kumar, JoydeepGhosh and • David J. Hand, “Top 10 algorithms in data mining”, Knowledge and Information System, PP. 1–37, (2008).
21. Vapnik V (1995), the nature of statistical learning theory. Springer, New York
22. Kavitha G, Udhayakumar A and Nagarajan D, “Stock Market Trend Analysis Using Hidden Markov Models”, available online: https://arxiv.org/ftp/arxiv/papers/1311/1311.4771.pdf.
23. Haiqin Yang, Laiwan Chan, and Irwin King, “Support Vector Machine Regression for Volatile Stock Market Prediction”, Intelligent Data Engineering and Automated Learning IDEAL, PP. 391- 396, Springer-Verlag Berlin Heidelberg 2002
24. PritamGundecha, Huan Liu, “Mining Social Media: A Brief Introduction”, 2012 INFORMS |isbn 978-0-9843378-3-5 http://dx.doi.org/10.1287/educ.1120.0105
25. AbeedSarker, Rachel Ginn, AzadehNikfarjam, Karen O’Connor, Karen Smith, SwethaJayaraman, Tejaswi Upadhaya, Graciela Gonzalez, “Utilizing social media data for pharmacovigilance: A review”, Journal of Biomedical Informatics, 2015 The Authors. Published by Elsevier Inc
26. Wu He, ShenghuaZha, Ling Li, “Social media competitive analysis and text mining: A case study in the pizza industry”, © 2013 Elsevier Ltd. All rights reserved.
27. Ido Guy, Uri Avraham, David Carmel, Sigalit Ur, Michal Jacovi, Inbal Ronen, “Mining Expertise and Interests from Social Media”, WWW 2013, May 13–17, 2013, Rio de Janiero, Brazil. ACM 978-1-4503-2035-1
28. H. K. Chan, E. Lacka, R. W. Y. Yee, M. K. Lim, “A Case Study on Mining Social Media Data”, 978-1-4799-6410-9/14/$31.00 ©2014 IEEE
29. Stefan Stieglitz, Linh Dang-Xuan, “Social media and political communication: a social media analytics framework”, Received: 2 February 2012 / Revised: 24 May 2012 / Accepted: 13 July 2012 Springer-Verlag 2012

30. Chaolun Xia, Raz Schwartz, KeXie, Adam Krebs, Andrew Langdon, Jeremy Ting, MorNaaman, "CityBeat: Real-time Social Media Visualization of Hyper-local City Data", WWW'14 Companion, April 7–11, 2014, Seoul, Korea. ACM 978-1-4503-2745
31. Hong-Han Shuai, Chih-YaShen, De-Nian Yang, Yi-FengLan, Wang-Chien Lee, Philip S. Yu, Ming-Syan Chen, "Mining Online Social Data for Detecting Social Network Mental Disorders", WWW 2016, April 11–15, 2016, Montréal, Québec, Canada. ACM 978-1-4503-4143-1/16/04.
32. Jun-Ki Min, Jason Wiese, Jason I. Hong, John Zimmerman, "Mining Smartphone Data to Classify Life-Facets of Social Relationships", CSCW '13, February 23–27, 2013, San Antonio, Texas, USA. Copyright 2013 ACM 978-1-4503-1331-5/13/02
33. Jiliang Tang, Huan Liu, "Unsupervised Feature Selection for Linked Social Media Data", KDD'12, August 12–16, 2012, Beijing, China Copyright 2012 ACM 978-1-4503-1462-6 /12/08
34. Kundan Kumar Mishra, Rahul Kaul, "Audit Trail Based on Process Mining and Log", International Journal of Recent Development in Engineering and Technology, Volume 1, Issue 1, Oct 2013
35. Rahul A. Patil, Prashant G. Ahire, Pramod. D. Patil, Avinash L. Golande, "A Modified Approach to Construct Decision Tree in Data Mining Classification", International Journal of Engineering and Innovative Technology (IJEIT), Volume 2, Issue 1, July 2012.