# A HYBRID TEXT CLASSIFIER FOR MINING SENTIMENTS ON TWITTER DATASET

Aaquib Multani [#1], Atul Agrawal[*2]

[#1]*M.E. Scholar, Sushila Devi Bansal College of Engineering, Indore, India.*

[*2]*Assistant Professor, Computer Science, Sushila Devi Bansal College of Engineering, Indore, India.*

*Abstract* – **Data mining techniques are used for analyzing the data and recovering the valuable information from the data. There two basic techniques of data analysis available namely supervised and unsupervised learning. The supervised algorithm are the accurate models by which the similar patterns are trained on the initial samples and accurate class labels can be identifiable for raw samples that not contains the class labels. In this presented work the supervised learning technique is used for classifying the unstructured twitter data for finding the emotional class labels. There are two class labels are available in data namely positive classes and negative classes. In order to perform classification the twits are first the pre-processed to make clean. In further the feature extraction technique is used by which the unstructured data is transformed into the structured format. For transformation of data NLP parser is used. The NLP parser is used to obtain the POS (part of speech) information from text and can be used for transforming the data into the 2D vector. This vector is used with the two different classification techniques namely C4.5 decision tree. First the decision tree model is developed using training data. And the developed decision tree is used for classify the test data. Further the data is again classified using the Bayesian classifier that replaces the misclassified data to improve the classification accuracy. The implemented technique is experimented a number of times and it is concluded the performance of the technique improves the classification ability by involving multiple classifiers.**

**Keywords:Data Mining, Sentiment Analysis, Twitter, Social Media, Multi-class Classification, Text Mining, Naïve Bayes.**

## I. INTRODUCTION

In this age of technology the internet and their service are common. The internet services are used in various applications such as banking, education and others. Among these services the social media is one of the popular applications. A significant amount of youth and students are expanding their time in the social media [1]. The social media provides the ability to share the information or data in this platform publically. In this platform when the users post their data from the text their emotions are also reflected. Therefore that text can be used for recognizing the moods of the end user [2]. In this presented work the aim is to classify the twitter data for finding the user behavior or mood.

The emotion classification or the sentiment based text classification is a new domain of research and development. But the nature of classification is complex due to the irregularity of the twit length and the text available. Therefore some new kind of technique is required that first transform the data into the structured data. The classification [3] of emotions required to include the text mining approaches and the NLP techniques to parse the data and classify them. In this context the process is detailed by which the classification is performed accurately [4]. The classification technique is a supervised algorithm for data analysis. The supervised learning algorithms first consume the initial data samples and then a data model is developed. This data model is a mathematical form of data or statically evaluation of data [5]. Additionally using the developed mathematical data model is used for further used for discovering the similar patterns appeared.

## II. PROPOSED WORK

The twitter data is needed to classify here for finding the sentiment class labels. By which the user's emotions are classified in terms of negative and positive patterns. This chapter contains the methodology and proposed algorithm for explaining the required system.

### A. System Overview

In data mining techniques the classification played an essential role. The classification is a supervised learning technique of data mining. In this technique some initial patterns (i.e. sample patterns or training samples) are used for developing the data model that is able to recognize the same pattern as training provided to the algorithm [6]. The training samples consist of the instance pattern and the predefined class label. During training the modeling is performed to identify the class labels and after training it is expected by the algorithm to recognize the similar pattern by predicting the class labels [7]. In this presented work the classification technique is used for predicting the sentiments of the text data patterns.

The text pattern is basically an unstructured data format. The unstructured data is not is similar by nature and by length. Therefore that is very complex work and need more effort to make data suitable to utilize with the classification algorithm. Therefore different pre-processing techniques need to apply for transforming the data into the structured data. The transformation of data is performed such that by which the suitable attributes or features form the unstructured data are make similar in length or nature for learning. Here the NLP (natural language processing) technique is used to transform the data into structured format. Finally the algorithm is trained for classifying the data. In this presented work a hybrid classifier is implemented for classifying the transformed data. The hybrid classifiers contains the goodness of the both the algorithm. In this work the decision tree algorithm is hybridized with the Bayesian classifier. The Bayesian classifier cross verify the classification performed which patterns are classified through the decision tree algorithm. This section provides the basic details about the proposed system. In next section the detailed system functions are explained.

### B. Methodology

The proposed data model which is used to recognize the emotion is reported in this section. The model demonstrated in figure 2.1 and the components of model are explained in the same section.

**Initial Samples:** the proposed work is to classify the twitter data for finding the sentiment classes through the text. Therefore a machine learning twitter data set is used here. The initial samples consist of the twits and the associated sentiment class in terms of negative and positive. The entire samples of this dataset are used for both the purpose i.e. training and testing of the proposed classification technique [8].

**Data Pre-Processing:** the initial twitter dataset is not in clean format. The twitter data contains different kinds of impurities such as unwanted characters and stop words. Therefore the provision is made to remove the unwanted characters and stop words from the twits. In order to remove both the unwanted data a function is implemented that accept the list of characters and stop words and replace the characters and stop words from the blank space [9].

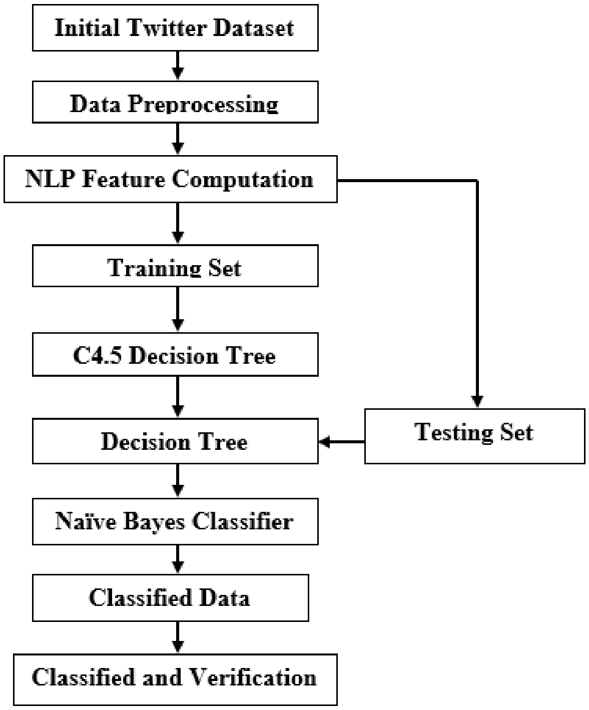

**Figure 2.1 Proposed System Architecture**

**NLP Feature Computation:** after cleaning the dataset the remaining data is not in such format by which the training and testing performed directly. Therefore the unstructured data format is need to be transformed into the structured format. In this context the Stanford NLP parser is used [10]. This parser is an open source JAVA API which can be used with the JAVA IDE as class library to utilize with the code lines. Using this parser the available twit instances are parsed into the part of speech information. This process of parsing the text into the part of speech information is termed as the POS tagging. After tagging of the data a 2D vector is prepared which is contains the attributes as the part of speech tags and the classes are defined with each instance of data. The table 2.1 contains example of the 2D vector of transformed twits.

| Noun | Pronoun | Verb | Adverb .... | Class label |
|---|---|---|---|---|
| 2 | 1 | 1 | 1 | Positive |

**Table 2.1 Example of POS Tagging**

The values of this 2D vector are the count of articles available in a single instance of twit.

**Training Set:** after transformation of the data unstructured formats to unstructured data format the two different set of data is prepared for performing training and testing of the classification algorithm. The 70% of randomly selected data is used as the training set of the system.

**Testing Set:** again form the entire dataset randomly 30% of randomly selected data instances is captured. That set of captured data is termed here as the testing set which is used for verification of trained classifier.

**Decision Tree C4.5:** C4.5 (Created by Quinlan, 1993) a calculation that takes in the choice tree classifiers, It has been watched that C4.5 performs short in the area where there is pre-passageway of consistent traits contrasted and the learning errands with generally isolate characteristics [11]. For example, a framework which searches for all around characterized choice tree with 2 levels and afterward put remarks:

"The precision of trees made with T2 is leveled or even surpass trees of C4.5 upon 8 out of all the datasets, with the whole aside from one that have unremitting qualities as it were."

Input: An exploratory informational collection of information (D) depicted with the methods for discrete factors.

Output: A choice tree say T which is built by methods for passing investigational informational collections.

1) A hub (X) is made;
2) Check if the occurrence falls in a similar class.
3) Make hub (X) as the leaf hub and dole out a name CLASS C;
4) Check IF the characteristic rundown is vacant, THEN
5) Make node(X) a leaf hub and dole out a name of most standard CLASS;
6) Now pick a characteristic which has most elevated data pick up from the gave property List, and afterward set apart as the test_attribute;
7) Confirming X in the part of the test_attribute;
8) In request to have a perceived an incentive for each test_attribute for partitioning the examples;
9) Generating a new twig of tree that is reasonable for test_attribute = atti from hub X;
10) Take a presumption that Bi is a gathering of test_attribute=atti in the examples;

11) Check If Bi is NULL, THEN

12) Next, include another leaf hub, with mark of the most broad class;

13) ELSE a leaf hub will be included and returned by the Generate_decision_tree.

**Developed Tree:** the C4.5 decision tree algorithm accepts the transformed twit data as the input and using this data the tree is constructed. The decision tree contains the nodes and edges in there development. The nodes of the decision tree contain the part of speech information articles and the edges which connect these edges contain the count of articles in a data instance. Finally in the leaf node of the decision tree contains the class labels of the classification.

**Classified Data:** after traversing the developed tree using the testing data instances. The class labels for each pattern are predicted. All the predicted classes are termed here the classified data through the decision tree.

**Bayesian Classifier:** The classical Bayesian classification theorem is described as:

The Naive Bayes classification algorithmic rule is a probabilistic classifier. It is based on probability models that incorporate robust independence assumptions. The independence assumptions usually don't have an effect on reality. So they're thought of as naive. You can derive probability models by using Bayes' theorem (proposed by Thomas Bayes). Based on the nature of the probability model, you'll train the Naive Bayes algorithm program in a very supervised learning setting. In straightforward terms, a naive Bayes classifier assumes that the value of a specific feature is unrelated to the presence or absence of the other feature, given the category variable. There are two types of probability as follows:

- Posterior Probability [P (H/X)]
- Prior Probability [P (H)]

Where, X is data tuple and H is some hypothesis. According to Baye's Theorem

$$P\left(\frac{H}{X}\right)=\frac{P\left(\frac{X}{H}\right)P(H)}{P(X)}$$

**Classified Data Verification:** that is the final classification system which accepts again the testing dataset instances and classified once again all the classified data which is used with the C4.5 decision tree algorithm. If the similar class labels are appeared after classification of a instance then system do nothing otherwise the class labels predicted by the C4.5 is changed to the predicted class according to Bayesian classifier predicted class label. After second time verification of classified class labels the performance of the system is measured in terms of accuracy and error rate.

### C. Proposed Algorithm

The above given entire process is concluded here as the algorithm steps. Table 2.2 shows the prepared algorithm steps:-

```
Input: training samples T
Output: predicted class labels C

Process:
1. $D_n = readDataset(T)$
2. $P_n = preProcessData(D_n)$
3. $for(i = 1; i \leq n; i++)$
       a. $POS_i = NLPParser.TagData(P_i)$
4. $endfor$
5. $[TestData, TrainData] = PartitionData(PSO_n)$
6. $TrainModel = C4.5.CreateTree(TrainData)$
7. $C = TrainModel.Classify(TestData)$
8. $C = Bayesian.ClassifyData(C)$
9. Return C
```

**Table 2.2 Proposed Algorithm**

## III. RESULT ANALYSIS

The given chapter provides the detailed understanding about the evaluated results of the proposed Multiclass Label Classification of social networks. Therefore this chapter includes the different performance parameters and their description on which the proposed system is evaluated using different size of data.

### A. Accuracy

In classification, accuracy is the measurement of accurately classified patterns over the total input patterns produced for classification result. Therefore this can be a measurement of successful training of the classification algorithm. The accuracy of the classifier can be evaluated using the following formula:

$$\text{Accuracy}=\frac{\text{Total correctly classified patterns}}{\text{Total input patterns to Classify}}X100$$

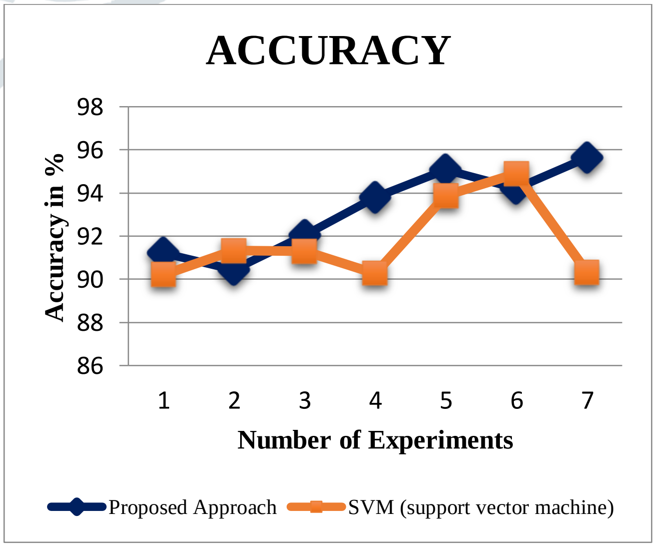

**Figure 3.1 Accuracy**

The accuracy of the implemented proposed algorithm and SVM classifier for classification of twitter dataset is represented using table 3.1 and figure 3.1. The given graph figure 3.1 contains the accuracy of the implemented algorithms. The X axis of the figure contains different experiments and Y axis contains the obtained performance in terms of accuracy percentage value. To demonstrate the performance of proposed sentiment analysis based tweet classification is represented dark blue line. And orange line describes the performance of classical SVM classifier. According to the obtained results the performance of the proposed model demonstrating multiclass classification of user tweets is most of the time higher than the SVM classifier. Additionally the accuracy of the feature classification model is increases as the amount of instances for the learning of algorithm is increases. On the other hand the performance of classical SVM classifier is fluctuating as compared to proposed hybrid classification technique.

**Table 3.1 Accuracy**

| Number of Experiments | Proposed Tweet Classification Approach | SVM (support vector machine) |
|---|---|---|
| 1 | 91.23 | 90.21 |
| 2 | 90.44 | 91.32 |
| 3 | 92.04 | 91.29 |
| 4 | 93.81 | 90.27 |
| 5 | 95.09 | 93.88 |
| 6 | 94.21 | 94.89 |
| 7 | 95.65 | 90.32 |

### B. Error Rate

The amount of data misclassified samples during classification of algorithms is known as error rate of the system. That can also be computed using the following formula.

$$\text{Error Rate} = 100 - \text{Accuracy}$$

**Table 3.2 Error Rate**

| Number of Experiments | Proposed Tweet Classification Approach | SVM (support vector machine) |
|---|---|---|
| 1 | 8.77 | 9.79 |
| 2 | 9.56 | 8.68 |
| 3 | 7.96 | 8.71 |
| 4 | 6.19 | 9.73 |
| 5 | 4.91 | 6.12 |
| 6 | 5.79 | 5.11 |
| 7 | 4.35 | 9.68 |

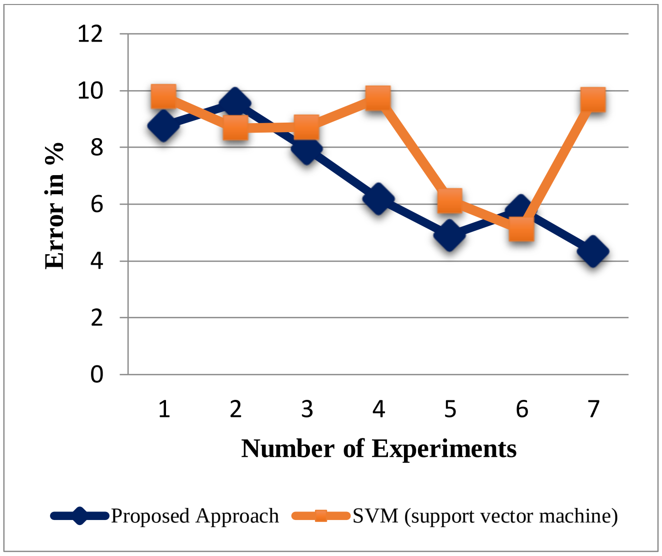

**Figure 3.2 Error Rate**

The figure 3.2 and table 3.2 shows the error rate of implemented both the classification algorithms. In order to show the performance of the system the X axis contains the performed different experiments and the Y axis shows the performance in terms of error rate percentage. The performance of the error rate of proposed naïve bayes and C4.5 classification is represented using blue line. Additionally the orange line shows the performance of traditional SVM classifier for text classification. The performance of the proposed classification is effective and efficient during different execution and reducing with the amount of data increases. On the other hand the performance of traditional SVM classifier is fluctuating with amount of data and noise contains. Thus the presented classifier is more efficient and accurate than the other implemented approaches of text classification.

### C. Memory Usage

Memory consumption of the system also termed as the space complexity in terms of algorithm performance. That can be calculated using the following formula:

$$\text{Memory Consumption} = \text{Total Memory} - \text{Free Memory}$$

**Table 3.3 Memory Consumption**

| Number of Experiments | Proposed Tweet Classification Approach | SVM (Support Vector Machine |
|---|---|---|
| 1 | 5625 | 6732 |
| 2 | 5059 | 6388 |
| 3 | 5561 | 6482 |
| 4 | 5980 | 6628 |
| 5 | 6223 | 6992 |
| 6 | 6624 | 7028 |
| 7 | 7002 | 7829 |

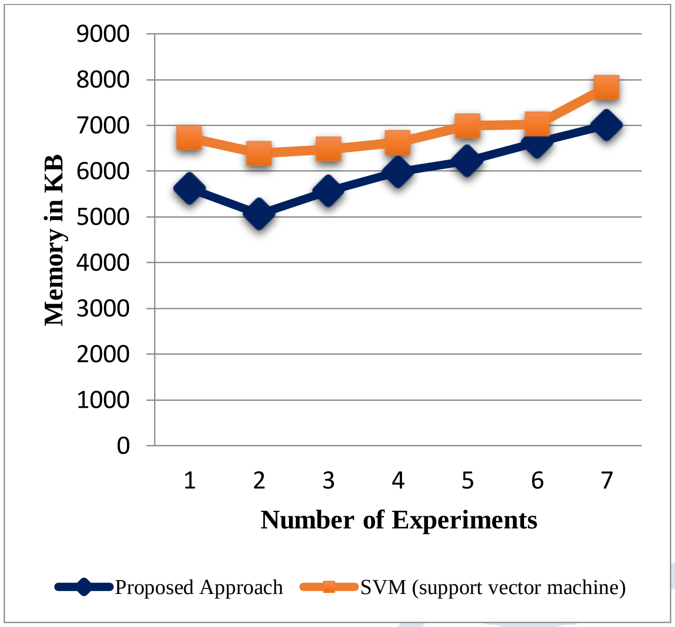

**Figure 3.3 Memory Consumption**

The amount of memory consumption depends on the amount of data reside in the main memory, therefore that affect the computational cost of an algorithm execution. The performance of the implemented both the classifiers for tweet classification are given using figure 3.3 and table 3.3. For reporting the performance the X axis of figure contains experiments and Y axis shows the respective memory consumption during execution in terms of kilobytes (KB). According to the obtained results the performance of algorithm demonstrates similar behavior with increasing size of data, but the amount of memory consumption is decreases with the amount of data. The SVM classifier needs more amount of main memory as compared to the proposed technique, because in one verses all scenarios the single class data is trained with respect to others therefore the SVM classifier needs additional amount of main memory as compared to the rule based classification technique.

### D. Time Consumption

The amount of time required to classify the entire test data is known as the time consumption. That can be computed using the following formula:

$$\text{Time Consumed} = \text{End Time} - \text{Start Time}$$

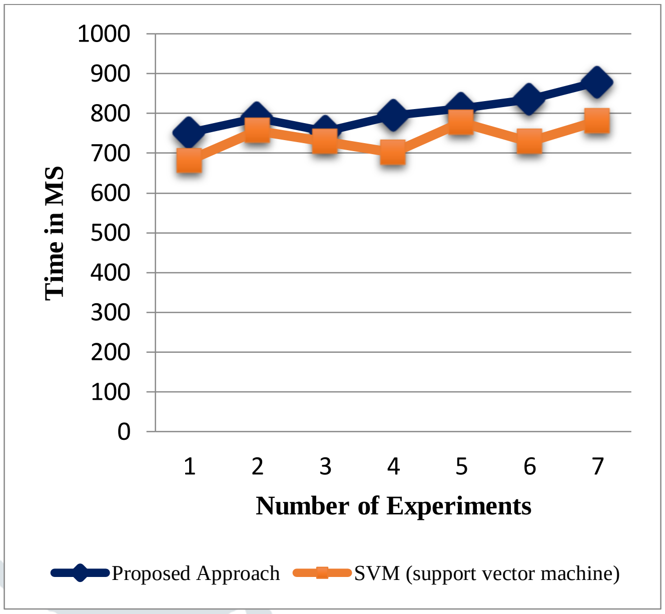

**Figure 3.4 Time Consumption**

The time consumption of the proposed algorithm and classical SVM classifier are given using figure 3.4 and table 3.4. In this diagram the X axis shows different experimentation and Y axis contains consumed time in terms of milliseconds. According to the results analysis the performance of the proposed technique minimize the time consumption. But the amount of time is increases in similar manner as the amount of data for analysis is increases. On the other hand for classifying similar amount of data the SVM classifier requires less amount of time. Thus in terms of classification time the SVM classifier is efficient as compared to the proposed data model due to rules and their evaluation time.

**Table 3.4 Time Consumption**

| Number of Experiments | Proposed Tweet Classification Approach | SVM (support vector machine) |
|---|---|---|
| 1 | 752 | 682 |
| 2 | 789 | 757 |
| 3 | 755 | 729 |
| 4 | 795 | 702 |
| 5 | 812 | 778 |
| 6 | 836 | 729 |
| 7 | 878 | 781 |

## IV. CONCLUSION

The main aim of the proposed work is to obtain the user's emotions or sentiments from the twitter data analysis. In this context a classification scheme is proposed and implemented successfully. This chapter contains the conclusion of the made effort based on the experiments and observations.

## A. Conclusion

In this age of technology almost all the people usages the service of the social media. Additionally the social media provides the faculty to write the post according to their sentiments and emotions. Basically when the author write something in social media, then the emotions are reflected in their expressed text. In this context by analysing the twitter or other social media data the author mood can be identified. In this presented work the twitter data is considered for classification and emotion class label prediction. The proposed classification technique double check the classification labels for ensuring the accurate classification of twitter data.

First the system accepts the data and pre-processes the entire data for removing the unwanted characters and stop words. After cleaning of the data transformation of the unstructured data is performed into the structured format. For that purpose the NLP (natural language parser) is used. The transformation of data converts the raw twit into the 2D vector that contains the attributes as the part of speech information and the class labels. Now first the C4.5 decision tree algorithm is applied for training with the data. After training of algorithm the test set of data is applied and the test dataset is classified using the developed decision tree. The classified data and predicted class labels the Bayesian classified is applied for verifying the classes predicted by the decision tree prediction.

The implementation of the proposed twitter data classification technique for finding the user's emotional conditions is performed using the JAVA technology. Additionally for holding the performance and intermediate structures of the data the MySql database is used. After implementation of the system the experiments on the implemented system is performed, additionally comparison of the performance is reported with respect to the traditional SVM classifier. The results are prepared on the basis of a number of times of experiments and based on the observations. The performance summary is prepared as reported in table 4.1.

| S. No. | Parameters | Proposed classifier | Traditional SVM |
|---|---|---|---|
| 1 | Accuracy | Higher and varying between (90-95 %) | Low and observed between (90-94 %) |
| 2 | Error rate | Low, it is observed between (5-10 %) | Higher it is found between (6-10 %) |
| 3 | Time consumption | Higher time requirements (780-880 MS) | Low classification time and found between (680-780 MS) |
| 4 | Space consumption | Low with respect to SVM found between (5000-7000 KB) | Higher then proposed technique observed between (6300 – 7800 KB) |

**Table 4.1 Performance Summary**

According to the obtained experimental results and the summary table as given in table 4.1 the proposed technique is acceptable for the real world data classification. In addition of that produces higher accuracy and low memory requirements as compared to the traditional SVM classifier. But time requirements of the system are higher which is needed to be improving in future.

## B. Future Work

The key objective of the proposed work is to enhance the classification ability of classifier to reduce the misclassification rate of unstructured data. Therefore a new model is implemented and designed. That mode is efficient and accurate and can be used for various other task. Based on the utility of the proposed classification technique the following future extension of the work is proposed.

1. The proposed technique extends the transparent data model for improving their classification performance by verification of classified data. In near future the ensemble learning technique is used for improving the performance more.
2. The proposed technique currently usage the transparent data model for classification and the transparent data models are less accurate as compared to the opaque data models. Therefore in near future the opaque data model is used for experimentation and system design

## REFERENCES

[1] Chen, Xin, Mihaela Vorvoreanu, and Krishna Madhavan, "Mining social media data for understanding students' learning experiences." IEEE Transactions on Learning Technologies 7, no. 3 (2014): 246-259.

[2] Chapter 3: Data Mining: an Overview, available online at: http://shodhganga.inflibnet.ac.in/bitstream/10603/11075/7/07_chapter3.pdf

[3] Mohammed J. Zaki and Wagner MeiraJr, "Data Mining and Analysis Fundamental Concepts and Algorithms", Cambridge University Press Hardback, 2014 [Book]

[4] Michael Goebel and Le Gruenwald ―A Survey of Data Mining and Knowledge Discovery Software Tools‖, ACM, 1999

[5] Neelam adhabPadhy, Dr. Pragnyaban Mishra, "The Survey of Data Mining Applications and Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), PP. 43-58 Vol.2, No.3, June 2012.

[6] Sundaravaradan, Naren, Manish Marwah, Amip Shah, and Naren Ramakrishnan. "Data mining approaches for life cycle assessment." In Sustainable Systems and Technology (ISSST), 2011 IEEE International Symposium on, pp. 1-6. IEEE, 2011.

[7] Gorunescu, F, Data Mining: Concepts, Models, and Techniques, Springer, 2011.

[8] Zhao, Yijun. "Data mining techniques." (2015).

[9] "Data Mining Tutorial: Process, Techniques, Tools & Examples", available online at: https://www.guru99.com/data-mining-tutorial.html

[10] N. Venkata Sailaja and L. Padmasree, "Survey of Text Mining Techniques, Challenges and their Applications", International Journal of Computer Applications (IJCA), Volume 146 – No.11, July 2016.

[11] Eman M.G. Younis, "Sentiment Analysis and Text Mining for Social Media Micro-blogs using Open Source Tools: An Empirical Study", International Journal of Computer Applications (IJCA), Volume 112 – No. 5, February 2015.

[12] Vishal Gupta and Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Volume 1, No. 1, PP. 60-76, August 2009.

[13] Bruno J. G. Praciano, João Paulo C. L. da Costa, João Paulo A. Maranhão, Fabio L. L. de Mendonça, Rafael T. de Sousa Junior, and Juliano B. Prettz, "Spatio-Temporal Trend Analysis of the Brazilian Elections based on Twitter Data", 2018 IEEE International Conference on Data Mining Workshops (ICDMW), 2375-9259/18/$31.00 ©2018 IEEE.

[14] Zhao Jianqiang, Gui Xiaolin, And Zhang Xuejun, "Deep Convolution Neural Networks for Twitter Sentiment Analysis", VOLUME 6, 2018, 2169-3536 2018 IEEE.

[15] Hajime Watanabe, Mondher Bouazizi, And Tomoaki Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection", VOLUME 6, 2018, 2169-3536 2018 IEEE.