Design and Implementation of Hierarchical Multi-Class Emotion Classification Model

Nidhi Jathar
CSE, RGPV, PG, Scholar PGOI Indore
M.P 452010/Indore, India

Abhilasha Vyas
CSE, RGPV, Reader PGOI, Indore
M.P 452010/Indore, India

ABSTRACT

Emotions play important role in human intelligence, Social Interaction, memory, learning, and more. Emotions are both prevalent in and essential to most aspects of our lives. With the rapid growth of emotion-rich textual content, such as microblog posts, Facebook posts, blogs posts, and forum discussions, such content can be used to unobtrusively identify and track people's emotions expressed in text. Social networks and micro-blogging tools such as Twitter allow individuals to express their opinions, feelings, and thoughts on a variety of topics in the form of short text messages. These short messages (commonly known as tweets) may also include the emotional states of individuals (such as happiness, anxiety, and depression) as well as the emotions of a larger group.In this research work, the sentiment is aimed to overcome the problem of automatically classifying user tweets into positive opinion and negative opinion. The classifier Naives Bayes (NB) used in this study is a machine learning technique that is popular text classifiers. Therefore, we proposed Multiclass Hierarchal Emotion based Classification using text mining applications to classify user tweets. Proposed method provides an effective way to immediately and accurately categorize multiclass sentiment tweets classification without need of exterior data, outperforming a content-based approach.The implementation of the proposed concept is provided using the JAVA environment. Additionally the comparative performance is also compared with traditional. In order to compare the performance of the algorithms the accuracy, error rate, memory consumption and time consumption is taken as standard parameters.

Keywords

SVM, Bayesian, Sentiment Analysis, Tweeter, Social Media, Classification, Text Mining, Multiclass Classification, POS

1. INTRODUCTION

Emotions are both prevalent in and essential to most aspects of our lives. They influence our decision-making, affect our social relationships and shape our daily behavior. With the rapid growth of emotion-rich textual content, such as microblog posts, blog posts, and forum discussions, there is a growing need to develop algorithms and techniques for identifying people's emotions expressed in text. In addition, micro-blogging sites are used as publishing platforms to create and consume content from sets of users with overlapping and disparate interests. As micro-blogging grows in popularity, services like Twitter are coming to support information gathering needs above and beyond their traditional roles as social networks.

The research work presents the text data for micro-blog analysis which is used for preparing the classification algorithm. Basically the micro-blogs are frequently used now in these days with small amount of communication data. But frequent use of this communication channel increases the amount of data for manual analysis. The presented system is a Multi class text analysis technique that works on the labeled data and provides the outcomes in two step process. In first the data is processed in order to extract the text features and then the learning on evaluated features is performed.

2. PROPOSED WORK

This chapter provides study about the concept of proposed multi-class emotion classification. In this context the different aspects of the solution formulation, there functional expectations and the required algorithm description is provided.

A. System Overview

The text mining is a part of data mining. In text mining more specifically the text data is treated for analysis and pattern discovery. The text mining techniques are become popular due to employment of the applications in various rich domains of applications. Analysis of digital documents, reviews about the products and services, digital library management are some essential domain of text data mining. In addition of that the text mining techniques are now in these days also used in emotion classification or the user sentiment analysis. Using such kind of emotion analysis we discover the moods of user, orientation about some kinds of service, obtaining the work load or stress level of end user. Therefore that is an essential topic of research and study.

In this presented work the emotion classification techniques using the text mining is investigated. Unlike the traditional sentiment analysis technique this technique not only classifies text data in two major classes positive or negative. That technique also suggest the useful for classifying data one more level, in other words the sub-classes of the negative and positive classes. Therefore the proposed model is termed as the heretical classification technique. Basically in order to deal with such kind of complicated text analysis task three phases of text mining process is applied. In first the preparation of training and testing data, secondly the nested Bayesian classification for training with the data additionally in this phase the NLP parser is also used for PSO tagging for effective data analysis, and finally the testing on randomly selected data for evaluation of effectiveness of the proposed text mining data model. This section provides the brief introduction of the proposed multiclass emotion classification technique using the text mining technique. The next section provides the detailed formulation of the proposed concept.

B. Methodology

This section provides the detailed understanding about the proposed multi-class text mining technique. As discussed in previous section the proposed model is defined in the major phases the descriptions of these phases are given as follows:

a. Dataset preparation

The data mining models are divided in two major parts supervised learning techniques and unsupervised learning technique. In supervised learning some initial input samples are required that are used for learn the patterns. In addition of that in unsupervised learning the models are directly perform evaluation on data. Traditionally the Bayesian is a supervised learning model therefore it required some predefined patterns to learn. In this context a heretical data set is prepared. In this dataset initially the two directories are prepared which are representing the two major classes of data. The figure 2.1 shows the initial dataset class labels.

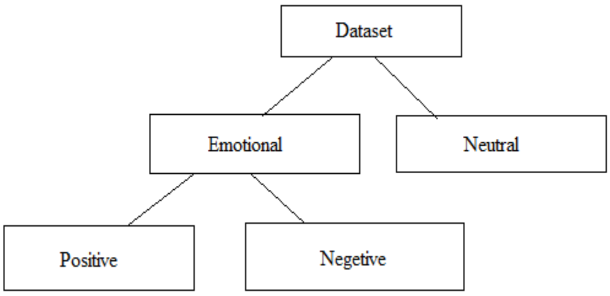

Figure 2.1 Dataset Organization

Initially the dataset is a root directory which contains the two sub-directories, namely emotional and neutral. The neutral class contains a file directly which contains the social media communicated or posted text. On the other hand the emotional directory again sub-divided in two labels positive and negative. Now these two directories contain the files with the social media text. In this text file the individual text post is again labelled with the following classes:

Positive:

- Fond
- Joyful

Negative:

- Distressed
- Surprized
- Fearful
- Angry
- Disgusted

This section described how the dataset for proposed data model is prepared as the training set, now for preparing the test set the same data with random selection process is combined in a separate file which is used for testing of the model.

b. Training Model

The proposed training model of the required system is given using figure 2.2, in addition of that their functional aspect is also described in this section. The different components of the system are demonstrated in this diagram that are processing the data and compute the input for next phase.

Read dataset: this process is responsible for loading of the dataset to the algorithm for further computing. As the data set is organized in heretical manner the reading and their class labels are also defining in this phase in the similar manner. Therefore an encoding scheme for processing of the data is presented which is usages the levels of data depth and utilized for constructing the class labels. The table 2.1 contains the process of the encoding for identifying the class labels of the data in their used heretical levels and according to their multiple context.

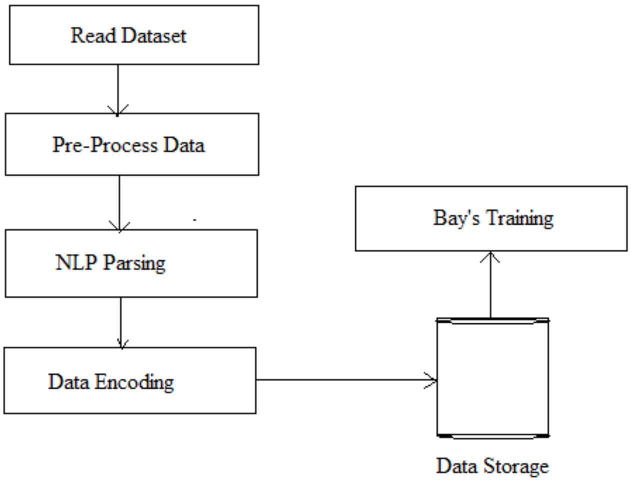

Figure 2.2 Training model

Table 2.1 dataset reading

```
Input: Dataset D
Output: text and classes C

Process:
    1. $R_n = ReadInitialDataset(D)$
    2. $for(i = 1; i \le n; i++)$
        a. $if(R_i == directory)$
            i. C.Add(DirectoryName);
            ii. Go to step 1
        b. Else
            i. $T = readFile(R_i)$
        c. $end\ if$
    3. End for
    4. Return C, T
```

Pre-processing: pre-processing is essential step of data mining. In this step the quality of data is enhanced. In addition of that in this step the cleaning of the data is also performed. Therefore low impact data or un-necessary data is removed from the original dataset. In this work the two different kinds of processes involved in first phase the special characters from the data is removed and in next the low impact data from the data set is removed. This low impact data is the basic words that are frequently used during the sentence formation such as the, is, am, are, this, that, so on. After cleaning of data, it is forwarded for the next phase.

NLP Parsing: NLP (natural language processing) is a very useful step for identifying the structure of the sentence. This provides the knowledge about the sentence part of speech. Thus this process is also termed as the part of speech tagging. In order to perform this task the NLP stand ford parser is used.

Data encoding: the structure or part of speech information is extracted from the previous phase and using this information the data is reorganized in the following manner as demonstrated in table 2.2.

Table 2.2 Data Encoding Example

Noun	Pronoun	verb	Adjective	Class
1	0	1	1	Angry
2	1	1	2	Joyful

Therefore according to the given table each sentence is parsed using the NLP parser and the frequencies or occurrence of the POS information a table is prepared which contains the class label also. Thus the sentences or the post on social media with their class labels are transformed in step from unstructured source of information to structured data.

Data storage: A data structure based temporary database is created which include the table which defined in previous step.

Bay's training: The standard approach to Bayesian classification uses the chain rule to decompose the joint distribution:

$$\Pr(C, A_1, A_2, \dots, A_k) = \Pr(C)\,\Pr(A_1, A_2, \dots, A_k|C) \dots\dots\dots\dots (1)$$

The first term on the right hand side of (1) is the prior probability of the class labels. These can be directly estimated from the training data, or from a larger sample of the population. For example, we can often get statistics on the number of, say, breast cancer occurrences in the general population. The second term on the right-hand side of (1) is the distribution of attribute values given the class label. The estimation of this term is usually more complex, and we elaborate on it below.

Once we have an estimate of $\Pr(C)$and $\Pr(A_1, A_2, \dots, A_k|C)$ we can use Bayes rule to get the conditional probability of the class given the attributes:

$$\Pr(C|A_1, A_2, \dots, A_k) = \alpha \Pr(C)\,\Pr(A_1, A_2, \dots, A_k|C) \dots\dots\dots\dots\dots (2)$$

where**α** is a normalization factor that ensures that the conditional probability of all possible class labels sums up to 1. (In practice, we do not need to explicitly evaluate this factor because it is constant for a given instance.) Using (2) we can classify new instances by combining the prior probability of each class with the probability of the given attribute values given that class.

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

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

c. Testing

This section described the testing of the proposed classification system. The figure 2.3 shows the testing model of the proposed system.

Testing dataset: the training dataset is used to create the test set of evaluation of classification model. Thus from all the files in random manner some post are selected and an additional file is prepared for utilizing as the test dataset.

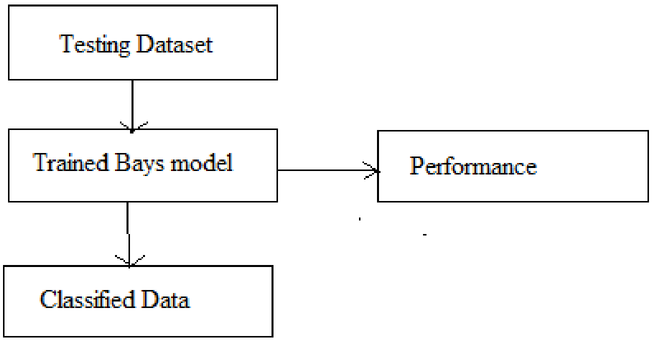

Figure 2.3 testing model

Trained bays model: in this phase the trained data model or computed probabilities of the words during the training is used for computing the class labels of the test dataset posts or text data.

Classified data: the computed class labels of the input test data is the classification of the test dataset.

Performance: during the evaluation of test dataset the according to the correctness of classification the performance of the algorithm is also computed which is used for further results analysis.

C. Proposed Algorithm

This section provides the combine algorithm for processing the input and their classification. The algorithm is given in table 2.3.

Table 2.3 proposed algorithm.

```
Input: training dataset D, test dataset T
Output: classified data C

Process:
    1. $R_n = readDataset(D)$
    2. $P_n = PreProcessData(R_n)$
    3. $for(i = 1; i \le n; i++)$
        a. $Tag_i = NLP.Parse(P_n)$
    4. $end\ for$
    5. $Tab_n = TransformData(Tag_n)$
    6. $Trained_{model} = Bays.MakeTraining(Tab_n)$
    7. $Ts_m = ReadTestDataset(T)$
    8. $for(j = 1; j \le m; j++)$
```

```
        a.   $C = Trained_{model}.Classify(Ts_j)$
9.   $end\ for$
10.  Return C
```

3. RESULTS ANALYSIS

The given chapter provides the detailed understanding about the evaluated performance of the proposed Multiclass hierarchal Emotion based Classification of social networks. Therefore this chapter includes the different performance parameters and description on which basis we compare base and proposed model.

A Accuracy

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

$$\text{Accuracy} = \frac{\text{Total Correctly Classified Patterns}}{\text{Total Input Patterns to Classify}} X100$$

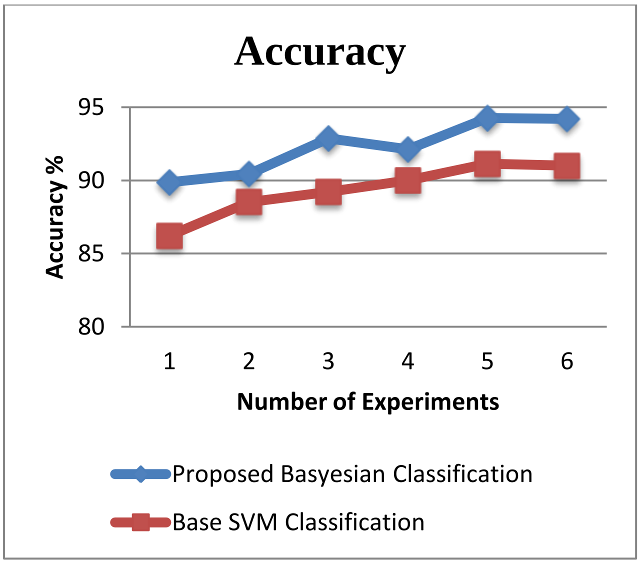

Figure 3.1 Accuracy

Table 3.1 Accuracy

Number of Experiments	Proposed Bayesian Classification	Base SVM Classification
1	89.861	86.214
2	90.441	88.521
3	92.875	89.203
4	92.112	91.002
5	94.258	91.335
6	94.201	91.002

The accuracy of the implemented proposed algorithm of emotion classification is represented using table 3.1 and figure 3.1. The given graph 3.1 contains the accuracy of the implemented algorithms. The X axis of the diagram contains the amount of data during the training and testing and Y axis contains the obtained performance in terms of percentage. To demonstrate the performance of both the techniques the blue line is used for proposed classification model and orange line shows the performance of SVM approach. According to the obtained performance the proposed model provides more accurate results as compared to base approach. Additionally the accuracy of the feature classification model is increases as the amount of instances for the learning is increase of algorithm.

B. Error Rate

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

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

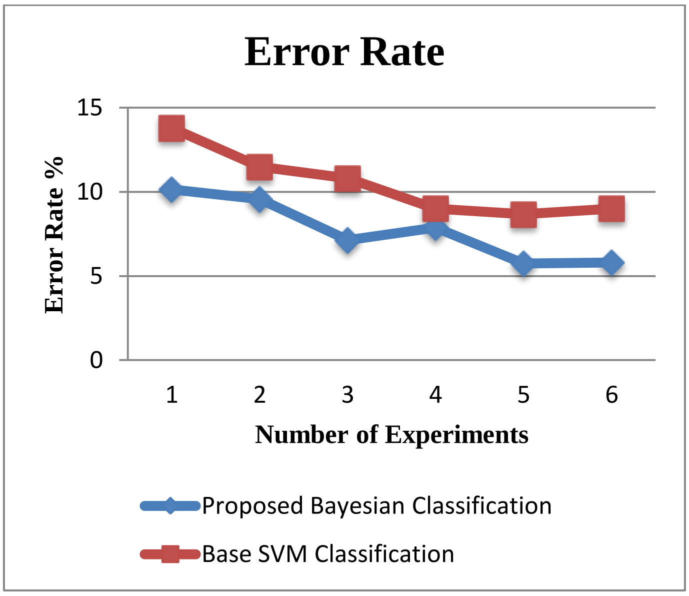

Figure 3.2 Error Rate

Table 5.2 Error Rate

Number of Experiments	Proposed Bayesian Classification	Base SVM Classification
1	10.139	13.786
2	9.559	11.479
3	7.125	10.797
4	7.88	8.998
5	5.742	8.665
6	5.799	8.998

The figure 3.2 and table 3.2 shows the comparative error rate performance of both implemented classifier. In order to show the performance, the X axis contains the amount of data used for training and the Y axis shows the performance in terms of error rate percentage. The error rate of the SVM method is given using the orange line and the performance of the proposed Bayesian classification technique is given using the blue line. The performance of the proposed classification is effective and efficient during different execution and reducing with the amount of data increases. Thus the presented classifier is more efficient and accurate than the traditional approaches of text classification.

C. Memory Usage

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

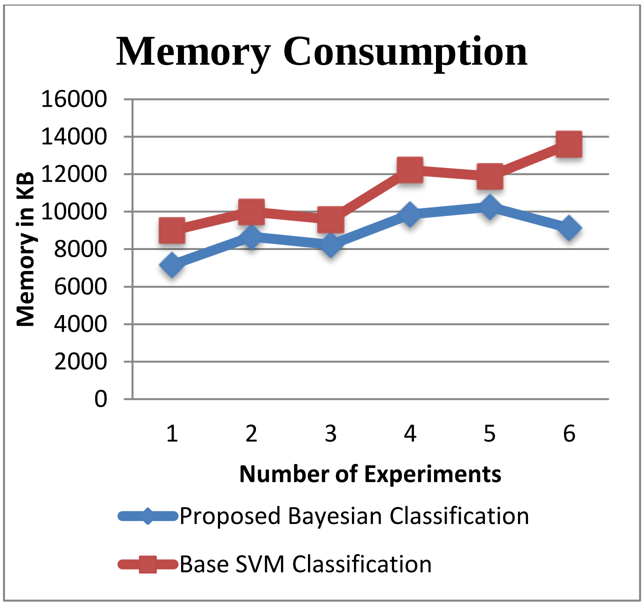

Figure 3.3 Memory Consumption

Table 3.3 Memory Consumption

Number of Experiments	Proposed Bayesian Classification	Base SVM Classification
1	7162	8994
2	8662	9989
3	8235	9584
4	9851	12220
5	10255	11889
6	9128	13605

$$Memory\ Consumption = Total\ Memory - Free\ Memory$$

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 classifier for multiclass hierarchal emotion classification is given using figure 3.3 and table 3.3. For reporting the performance the X axis of figure contains the amount of data required to execute using the algorithms and the Y axis shows the respective memory consumption during execution in terms of kilobytes (KB). According to the achieved performance the algorithm demonstrates similar behavior while we executing the system repeated, but the amount of memory consumption is decreases with the amount of data.

D. Time Consumption

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

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

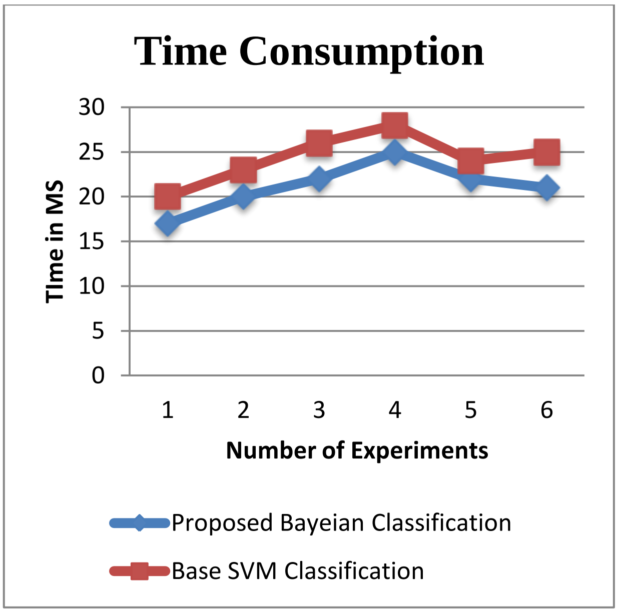

Figure 3.4 Time Consumption

Table 3.4 Time Consumption

Number of Experiments	Proposed Bayesian Classification	Base SVM Classification
1	17	20
2	20	23
3	22	26
4	25	28
5	22	24
6	21	25

The time consumption of the proposed algorithm is given using figure 3.4 and table 3.4. In this diagram the X axis contains the size of dataset and the Y axis contains time consumed in terms of milliseconds. According to the comparative 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.

4. CONCLUSION AND FUTURE WORK

The presented work is elaborated to find trained data mining approaches which are used to classify the text data according to user's emotions. Thus, this study has focused on analyzing the text mining methods and the classification algorithms. This chapter provides the summary of whole proposed work to performed user text classification. Additionally, in this chapter brief future extensions are also listed.

A. Conclusion

The performance of the system is estimated for finding the system accuracy and error rate for the multiclass hierarchal classification. Additionally for performance and efficiency the time and space complexity of the system is evaluated. The performance summary of system is given using table 4.1.

Table 4.1 Performance Summary

S. No.	Parameters	Proposed Bayesian Classification	Base SVM Classification
1	Accuracy	High	Low
2	Error rate	Low	High
3	Memory	Low	High
4	Time	Low	High

According to the obtained results the system is able to classify the data according to their classification and frequency accurately. Thus the proposed model for text classification is desirable and efficient. This work makes empirical contributions to this research area by comparing the performance of different popular sentiment classification approaches and developed collaborative approach, which further improves the sentiment classification performance.

B. Future Work

The key aim of the proposed work is to design and implement the multiclass text classification for finding the emotions expressed in text is accomplished successfully. In near future the following extensions are considered for the work.

1. Implementation of the system with a real world application for analyzing the feedback and reviews for a product or service
2. Investigate about more classification techniques that are improve the current classification performance.

5. REFERENCES

[1] Xu, Hua, Weiwei Yang, and Jiushuo Wang. "Hierarchical emotion classification and emotion component analysis on Chinese micro-blog posts." Expert systems with applications 42.22 (2015): 8745-8752.

[2] Mining, What is Data, "Data Mining: Concepts and Techniques." Morgan Kaufinann (2006).

[3] Han, J., and Kamber, M., Data mining: Concepts and techniques, Morgan-Kaufman Series of Data Management Systems San Diego: Academic Press, 2001.

[4] Neelam adhab Padhy, Dr. Pragnyaban Mishra and Rasmita Panigrahi, "The Survey of Data Mining Applications and Feature Scope, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)", vol.2, no.3, June.

[5] Dheeraj Agrawal, "A Comprehensive Study of Data Mining and Application", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 2, Issue 1, January 2013.

[6] Delmater R and Hancock M, Data Mining explained-a manager's guide to customer-centric business intelligence (Digital Press, Boston) 2002.

[7] Aakanksha Bhatnagar, Shweta P. Jadye, Madan Mohan Nagar" Data Mining Techniques & Distinct Applications: A Literature Review" International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 9, November- 2012

[8] Industry Application of data mining, available online at: http://www.pearsonhighered.com/samplechapter/0130862711.pdf

[9] Jiban K Pal, "Usefulness and applications of data-mining in extracting information from different perspective", Annals of Library and Information Studies, Vol-58, March 2011, pp. 7-16.

[10] Miss Latika Kaushik, "Text Mining - Scope and Applications", Journal of Computer Science and Applications, Volume 5, Number 2 (2013), pp. 51-55

[11] A.H. Tan, Text Mining: The State of the Art and the Challenges, in PAKDD99 Workshop on Knowledge Discovery from advanced Databases, Beijing, China, April 1999.

[12] Nahm U.Y. e Mooney R.J., Using Information Extraction to Aid the Discovery of Prediction Rules from Text, in KDD2000 Workshop on Text Mining, Boston, Massachusetts, USA, and August 2000.

[13] Dr. S. Vijayarani Ms. J. I lamathi and Ms. Nithya, "Preprocessing Techniques for Text Mining - An Overview", International Journal of Computer Science & Communication Networks, Volume 5(1), pp. 7-16

[14] Vishal Gupta, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009

[15] Li, J. & Khan, S. U. 2009. MobiSN: Semantics-based mobile ad hoc social network framework, In Proceedings of IEEE Global Communications Conference (Globecom), Honolulu, HI, USA.