Twitter Text Objectionable Content Detection using Domain Based Probability and Correlation Model

Divya Goyal[1], Dr. Pramod S Nair[2], D.Srinivasa Rao[3]

[1]*Research Scholar, Department of CSE, MITM, Indore, M.P., India*
[2]*Professor, Department of CSE, MITM, Indore, M.P., India*
[3]*Associate Professor, Department of CSE, MITM, Indore, M.P., India*

***Abstract*— Text mining technique is used to detect the required patterns from the text documents. The nature of text document is an unstructured format where the data is represented in an unstructured manner. The text mining can be used for information retrieval, information extraction, search, classification, and categorization. In this context, an application of text mining is proposed in this work. That effectively analyzes the context of word utilization and provides their context as class label. The proposed work is a model of text classification for detection of illegal use of words in text communication. Thus the proposed technique works in two modules first it trained with the different context of the text and then uses the features to classify the upcoming text as testing. During training, the word probability and the word's domain wise probability is estimated. Additionally, this information keeps preserved in a database for testing purpose. In the next, phase the testing of the system initiated through the training database and a test set supplied by the experimenter. During this process, all the sentences in a testing datasets are evaluated for computing the sentence probability and correlation estimation. Further, both the parameters are used to compute the weights. These weights are converted into a different indicator named as weight transform. Finally, a threshold is computed for making a decision. The proposed objectionable content detection technique using probability model and correlation is developed using JAVA environment. The implemented model is evaluated and compared with respect to their classical version of objectionable content detection. Results show the improvement made on traditional work improves their ability in terms of accuracy. Thus the model is acceptable for real world applications too.**

***Keywords*- text mining; content detection; objectionable content detection; pattern matching; text classification**

I. INTRODUCTION

Text mining techniques are used in various applications for discovering the valuable patterns [1]. These applications are not only used for categorizing and classifying the content but also used in various other applications such as terror attack detection, user's sentiment analysis, user's review about products [2] and services [3]. In this work, the text mining technique is studied in order to find the objectionable contents of the text communication. Basically, in text mining, the data mining techniques are used to their basic functionality, but before processing the data it is required to be transformed and to be converted to such a format by which algorithm can accept the data and process it.

The basic idea of the proposed work is taken from the web content analysis technique, where the web page contents are mined in order to find the required patterns [4]. But in this work the communicated text in web such as social networking sites, micro-blogs are intended to mine for finding the objectionable content in web data. The concept behind the proposed text mining technique is to identify such content from the text that is semantically objectionable to use in normal communication. Therefore it is necessary to learn about the use of the words in different situations and different context. In order to perform this task, a data model is required that help to understand the utilization of the words in different situations. In this context, the probability based models are much appropriate to find the patterns from data and compare them to detect for finding the required classes among the available mixed content. Thus the probabilistic model helps to classify the contents in two major classes objectionable and non-objectionable. In this section, the basic concept of the proposed work is provided and in the next section the key objective, motivational article and the background of text content mining is provided.

II. PROPOSED WORK

This chapter provides the detailed understanding of the proposed methodology for computing the web objectionable content. In order to perform such task, the text mining technique is included and using this technique the required pattern of data is extracted or identified.

A. System Overview

In real world practice, a word may be used in various places and context. Additionally every place of using the word, their meaning, semantic and sentiment may differ from each other. In some conditions, a word becomes cheerful or normal in use and in some of the places it understood as objectionable. In this presented work the main aim of the work is to identify a word when becomes objectionable and when it is used as legitimate in manners. In order to obtain this goal need to analyze the context of words. Therefore a word which is used in different kinds of context and contents are needed to be understood and then we can use the difference of utilization

for computing the new coming data [5]. For performing such complex data analysis some new kind of data modeling is required which is inherited from the text mining and probability theory.

In order to accomplish the objective, the three step process is incorporated in this work. In the first step, the data is pre-processed to refine and improve the quality of data. In the second phase, the learning is performed to learn the pattern of data as specified in different kinds of data set. In further, the threshold-based approach is included to compute the context of word utilization. Thus the probability based model is needed to compute the word utilization probability in different conditions. This obtained probability distribution makes enable to compare the obtained values and find the decision of content utilization. To improve the probability distribution's strength the correlation coefficient is also computed and using both the parameters the final decision is made. This section provides the overview of the proposed technique and the next section provide the details about the proposed model.

B. Proposed Methodology

The proposed model for objectionable content detection is reported in this section. The entire work is subdivided into two major modules. In first, the input data is processed for computing the probability of the data set that is termed as training phase of the system. In next phase, the data which is needed to be classified is provided as input and the objectionable contents or utilization of word is detected.

a. Training Module

The training module of the proposed system is described using the figure 2.1. In this diagram the participating components are described as follow:

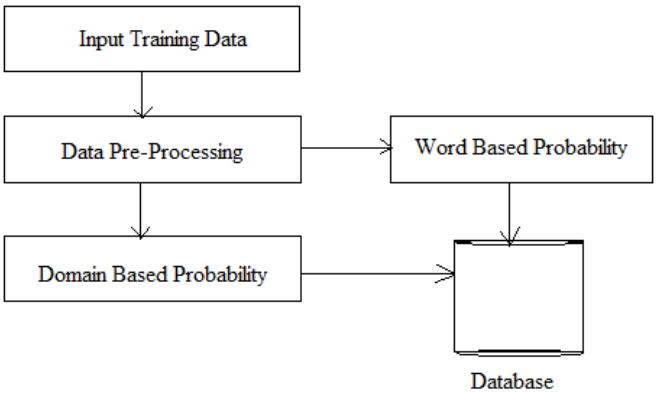

Figure 2.1 Training Model

Input Training Data: The main aim of the work is to develop the technique for finding the objectionable content according to the context of utilized term or word. Therefore it is required to define a learning dataset which is that is composed of the different aspects of the word utilization. In order to perform or develop a training dataset, three different contexts of similar words are defined the details about the word and utilization is demonstrated as the training dataset.

1. **Full objectionable sentence:** This is a set of data or text files that include the fully objectionable content and the utilization of word. That helps to identify the context of words how these words become fully objectionable.
2. **Social networking site:** In this set of text the socially used context of the word needs to be found. Thus the data from the social networking site communication is extracted; here the Twitter dataset is used.
3. **Third Google search content:** In this set of words or text we are considering the text that is normally or legitimately utilization of words. Thus the similar objectionable word is used to search on Google and the relevant record omitted is collected as the text set.

All the collected data is stored in different files and the used for preparing the training dataset which is further used for learning the pattern on the basis of the training dataset.

Data Pre-processing: The collected data is needed to be refined for obtaining the data which is significant for objectionable content detection. Therefore all the unnecessary data from the training dataset is removed. So the two processes are involved in reducing the unwanted content.

1. **Removal of special characters:** In this phase, the find and replace function is used for removing the special characters from the input text.
2. **Removal of stop words:** In this phase, the stop words from the data are reduced. The stop words are those words that are not having much significant for identification of any domain or subject such as objectionable or legitimate.

Domain-Based Probability: The proposed system involves the computation of two different factors for finding the accurate classification of data according to the probability distribution of words. That the first and essential parameter, in this probability, each participating word is evaluated with respect to their availability in different aspects or context. This probability is measured in the following manner:

$$PD_w = \frac{Total\ occerence\ of\ word\ in\ a\ domain}{Total\ words\ in\ domain}$$

Word-Based Probability: After computing the probability on the basis of domain or subject the word probability is computed for the entire dataset. That is termed as the word based probability. That is computed in the following manner:

$$P_w = \frac{Total\ occerence\ of\ word}{Total\ words}$$

Database: A database is prepared for storing the data factors recovered from the input training set. Previously two factors PD_w (domain based word probability) and P_w word probability are computed and stored in a database. That is termed as the training database.

b. Testing Module

The testing module of the system is described using figure 2.2. The required components and their functional aspects are described in detail as follows:

Training Database: This is a database which prepared in the previous phase of learning, where two different parameters PD_w (domain based word probability) and P_w (word probability) is computed. This parameter is used as the primary parameter for the testing module.

Test Dataset: it is the second input parameter for the system processing and the data execution. A set of randomly selected data from the initial training set is created from the considered the entire domain. Additionally, it is considered the model is helps to classify the data according to the objectionable context of used words and legitimately used context.

Pre-process Data: As the training dataset is processed for improving the quality of data in training phase, we consume the similar data in this phase too. Therefore it is required to apply both the processes namely stop word removal and special character removal in this phase also for improving or extraction of meaningful words from the data.

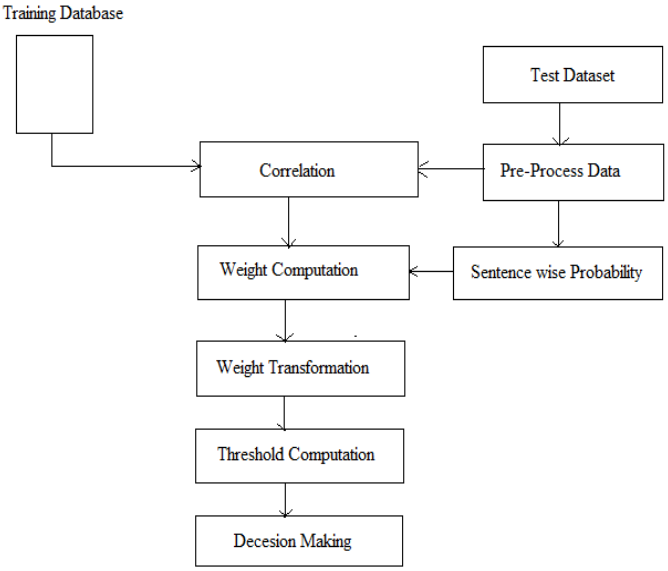

Figure 2.2 Testing Module

Sentence Wise Probability: the pre-processed data is used for computing the probability of sentence. Therefore all the participating word in each testing sentence is evaluated for all the three domains and their combined probability is computed.

Correlation: on the other hand the probability available in the database for the particular words in training database is used with the sentence word probability and their correlation is computed using the formula given below, the computed correlation return a value for each sentence, that is denoted using r:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i - y)^2}}$$

Weight Computation: The probability of sentence and the correlation factor is used to compute a weight for the sentences. In order to compute the weight using both the parameters namely sentence probability S_p and the correlation r, the following formula is used.

$$W = w_1 * S_p + w_2 * r$$

Where the $w_1\ and\ w_2$ is user defined values but need to satisfy the $w_1 + w_2 = 1$ condition. In this work, the values of $w_1\ and\ w_2$ is remains fixed 0.5.

Weight Transformation: The computed weight is used in this phase to transform the weight to prepare the new decision indicator.

$$nw = w * mf(ra)$$

Where mf is a mapping function that can be defined as follows:

$$mf(ra) = e^{1-\lambda * ra}$$

Where the λ is a factor which is also taken as constant and supposed 1.4. And ra can be defined as:

$$ra = \frac{rc_1}{rc_1 + rc_2}$$

Where $rc_1 and\ rc_2$ denotes the number of words which are in model word space and satisfies the constraints $0 \leq ra \leq 1$

Threshold Computation: In order to make a decision the weight threshold the mean weight value is used and computed as follows:

$$t = \frac{1}{N}\sum_{i=1}^{N} W_i$$

Decision-Making: In this phase, the obtained amount of weight from the computed threshold is compared and their orientation is obtained in terms of objectionable context or non-objectionable context.

C. Proposed Algorithm

This section provides the understanding about the proposed algorithm steps. As described in above section the proposed model is developed in two major phases training and testing. Thus for both the processes the two algorithms are prepared.

Input: Training Dataset D Output: training database TD
Process: 1. $R = readTrainingData(D)$ 2. $[OB_1, OB_2, OB_3] = R.getObservation(R)$ 3. $for(i = 1; i \leq 3; i++)$ a. $PD_w = DomainwiseProbebility(OB_i)$ b. $P_w = WordProbebility(OB_i)$ c. $TD.ADD(word, OB_i, PD_w, P_w)$ 4. $end\ for$ 5. Return TD

Table 2.1 Training Algorithm

Input: Training database TD, Test dataset TS Output: class labels {Obj, Nobj}
Process: 1. $R_{ts} = ReadTestset(TS)$ 2. $for(i = 1; i \leq R_{ts}.length; i++)$ a. $S = R_{ts}^{i}$ b. $for(j = 1; j \leq S.words; j++)$ i. $sp = SentanceProbebility(S, TD)$ ii. $r = Correlation(sp, TD)$ iii. $W = sp * w_1 + r * w_2$ c. $end\ for$ d. $nw = weightTransform(W)$ e. $t = \frac{1}{N}\sum_{i=1}^{N} W_i$ f. $if(S.weight > t)$ i. Result=Assign label Obj g. Else i. Result=Nobj h. End if 3. End for 4. Return Result

Table 2.2 Testing Algorithm

III. RESULTS ANALYSIS

This chapter provides the understanding about the performed experiments and obtained results after execution of both the implemented approaches of the web objectionable content detection system. The used parameters and their corresponding outcomes are described in this unit.

A. Classification Accuracy

The classification accuracy is the performance measurement of a classifier in order to indicate how accurately the classification is performed. Therefore the accuracy is the ratio of total correctly classified patterns over total patterns given for classifying.

$$Accuracy = \frac{Total\ correctly\ classified\ pattern}{Total\ pattern\ to\ classify}$$

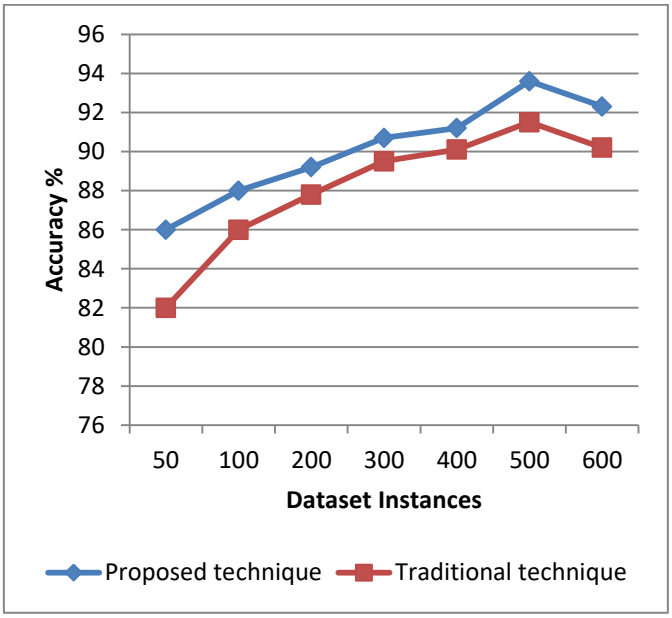

Figure 3.1 Comparative Accuracy

Dataset instances	Proposed technique	Traditional technique
50	86	82
100	88	86
200	89.2	87.8
300	90.7	89.5
400	91.2	90.1
500	93.6	91.5
600	92.3	90.2

Table 3.1 Comparative Accuracy

The comparative performance of proposed and traditional in terms of accuracy is given in table 3.1 and table 3.1. The measurement of accuracy is provided in terms of percentage values. In the given diagram blue line indicates the

performance of offered approach by us and the red line provides the performance technique of traditional topic-based objectionable content detection model. To represent visually X axis contains the size of data instances in experiments and the corresponding accuracy is described using Y axis. According to the experimental results both the technique simulates the similar behavior and improves the accuracy with the size of data. But after the number of instances 500 it becomes fluctuating, therefore, the performance is obtained between 88-94% in proposed technique. Thus the performance of the proposed technique is efficient than the traditional technique of objectionable content detection.

B. Error Rate

The error rate is the parameter of incorrectness of the data mining analysis process. That can also define as the amount of incorrectly identified samples over the total samples produced for identification is termed as the error rate of an algorithm.

$$Error\ Rate = \frac{Incorrectly\ identified\ pattern}{Total\ patterns\ to\ identify} X100$$

Or

$$Error\ Rate = 100 - Accuracy$$

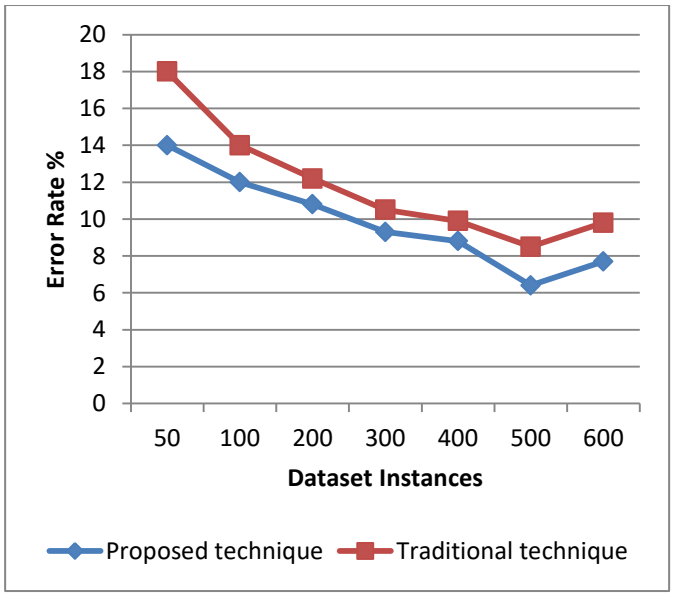

Figure 3.2 Comparative Error Rate

Dataset instances	Proposed technique	Traditional technique
50	14	18
100	12	14
200	10.8	12.2
300	9.3	10.5
400	8.8	9.9
500	6.4	8.5
600	7.7	9.8

Table 3.2 Comparative Error Rate

The error rate of both the implemented algorithm is demonstrated using table 3.2 and figure 3.2. The performance of offered approach is given using a blue line graph and the red line graph shows the performance of classical topic model based technique. In the similar diagram, the X axis represents the amount of data instances used for experiments and the obtained percentage error rate according to the increasing amount of data is defined in the Y axis. The experimental results show the proposed technique produces less amount of error as compared to the previous approach of objectionable content detection. Therefore the proposed technique is acceptable as compared to traditional approach.

C. Memory Usages

The requirement of main memory for execution of any algorithm or group of instructions is known as memory usages or space complexity of the algorithm. The figure 3.3 and table 3.3 shows the memory requirements of both the algorithms. The X axis of the diagram shows the amount of data used for experimentation and according to the experimental data, the required amount of main memory is given in Y axis. During the performance evaluation, the performance of algorithms is measured in terms of kilobytes (KB). According to the result,s the proposed technique requires higher memory resource as compared to the traditional technique because the proposed technique includes the computation of correlation also for improving the likelihood of classification. Thus this produces the additional impact on memory utilization.

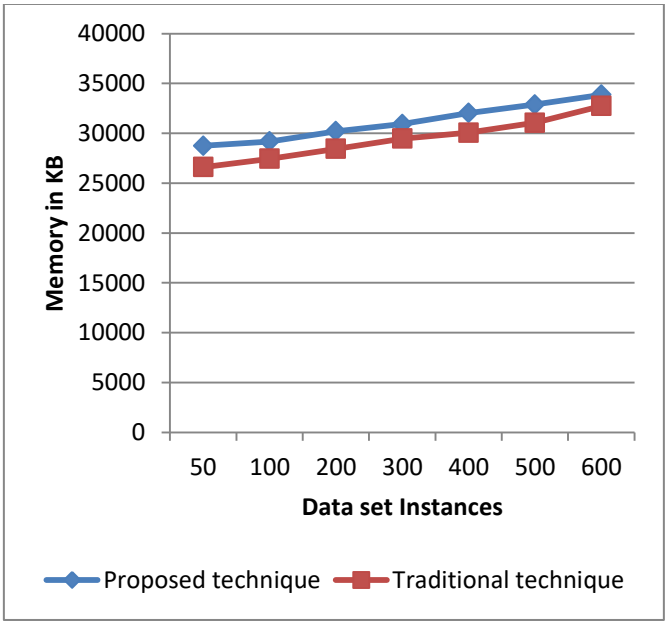

Figure 3.3 Memory Usage

Dataset instances	Proposed technique	Traditional technique
50	28745	26611

100	29184	27451
200	30197	28442
300	30918	29461
400	32041	30058
500	32890	31047
600	33857	32746

Table 3.3 Comparative Error Rate

D. Time Requirements

The amount of time required to analyze the data using an implemented data mining algorithm is termed here as the time requirement of the system. That is basically the amount of time difference between initiation of data analysis and completion of data analysis. Figure 3.4 and Table 3.4 is used to demonstrate the comparative performance reporting between both the algorithms. The X axis of diagram contains the size of experimental data in terms of number of instances.

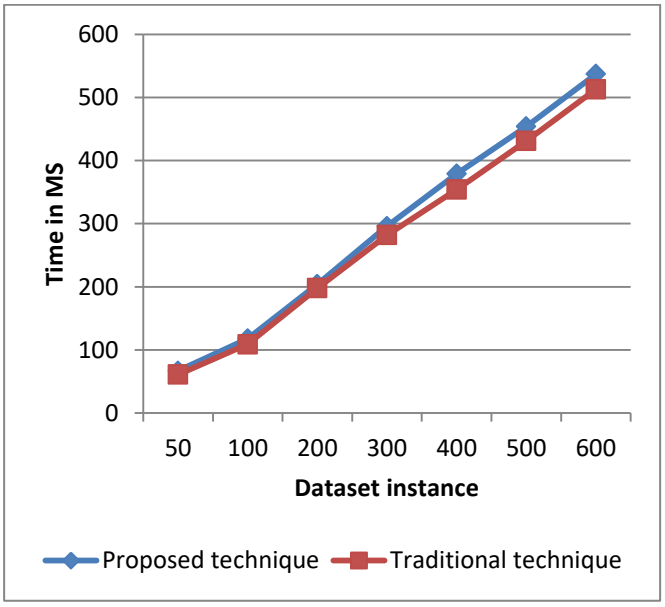

Figure 3.4 Comparative Time Requirements

Dataset instances	Proposed technique	Traditional technique
50	67	61
100	118	109
200	204	198
300	296	282
400	379	354
500	454	431
600	537	513

Table 3.4 Comparative Time Requirements

Additionally, the Y axis includes the time requirements for data analysis according to the input size of data. The time requirement of data is measured in terms of milliseconds. According to the computed results, the proposed technique requires more time as compared to the traditional technique because the proposed technique includes the computation of correlation parameter additionally. Thus the proposed technique is acceptable with the given performance.

IV. CONCLUSION AND FUTURE WORK

This chapter draws the conclusion of the performed study. The conclusion is made on the basis of experimentation and on the basis of an observation made during the design process. In addition of that, the feasible future extension of the work is also included in this chapter.

A. Conclusion

The data mining techniques are applied to different kinds of data for analyzing them using the computational algorithms when these techniques are applied on the text that is termed as the text mining technique. The text mining can be utilized for discovering the patterns, identification of similar patterns and analyzing the contents of input data. In this work, the web contents are analyzed for finding the objectionable content in a text communication. Thus in order to experimentation and design of an effective model, the twitter text dataset is considered. In addition of that, the Google search contents are also included for experimentation. Therefore it is required to estimate the context of words where it becomes objectionable and when it is legitimate for use. Thus in order to find such kinds of pattern in data, it is required to involve the context learning and semantic pattern mining technique is helpful. Therefore a probability-based model is proposed for investigation and design. Additionally, for enhancing the classification ability the correlation of context of a word is also included to work.

The proposed technique consists of two main phases of execution training and testing. During the training, three different sources of data are taken and pre-processing applied on data. The pre-processing technique is used to refine the data; additionally, the weighted contents or only essential words are remaining for further utilization. In next the remaining data is used to compute the domain based probability distribution of the word and token based word probability. This computation is extended for computing the weights of the words using probability distribution and correlation computation. After that testing of the system is performed therefore a mixed dataset is prepared and after pre-processing of test dataset the probabilities of data is computed. Finally, the weight transform is performed and a threshold value is also computed. This threshold value is helps to find which word in objectionable in which context.

The implementation of both the techniques namely proposed correlation based and classical topic model based technique is performed using JAVA technology. After implementation, the performance of the algorithms is computed and compared with

different parameters. The aim of evaluations of these parameters is to find the effectiveness and their efficiency to work with the text data. The obtained performance is described using Table 4.1. The table contains the range of values between which the performance varies.

S. No.	Parameters	Proposed technique	Topic model
1	Accuracy	86-92.3 %	82-90.2 %
2	Error rate	14-7.7 %	18.9.8 %
3	Memory	28745-33857 KB	26611-32746 KB
4	Time	67-537 MS	61-513 MS

Table 4.1 Performance Summary

According to the experimental summarization, as demonstrated in Table 4.1, the proposed technique is accurate but includes overhead for time and memory resource. But the obtained overhead is acceptable according to obtained accurate outcomes. Thus the proposed aim is accomplished according to the requirement and objective of the proposed work. This system can help in various real world applications i.e. terror, porn and other kinds of text communication.

B. Future Work

The proposed work is focused on improving the existing technique of web objectionable content identification, which is completed successfully. In near future, the work is extended in the following feasible directions.

1. The present work is used with the Twitter text dataset, in near future the micro-blog and live web data is used for experimentation and system design
2. The current work is modified for adopting other data formats to make it more effective for the objectionable content detection such as image, audio, and video

REFERENCES

[1] Vishal Gupta and Gurpreet S. Lehal, “A Survey of Text Mining Techniques and Applications” Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009.
[2] Delmater R and Hancock M, Data Mining explained-a manager’s guide to customer-centric business intelligence (Digital Press, Boston) 2002.
[3] Tan, Ah-Hwee, "Text mining: The state of the art and the challenges", Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, Volume 8, 1999.
[4] Duan, Jiangjiao, and Jianping Zeng, "Web objectionable text content detection using topic modeling technique", Expert Systems with Applications 40.15 (2013): 6094-6104.
[5] D. Jurafsky and J. H. Martin, Speech and Language Processing: An introduction to Natural language Processing, Computational Linguistics and Speech Recognition. United States of America: Prentice Hall, 2009
[6] Amrut M. Jadhav and Devendra P. Gadekar, “A Survey on Text Mining and Its Techniques”, International Journal of Science and Research (IJSR), Volume 3 Issue 11, November 2014