ISSN:0975-9646

# A System to Filter Unwanted Words Using Blacklists In Social Networks

K.Babu , P.Charles

*Department of Computer Science,*
*MRK Institute of Technology*
*Kattumannarkoil-608 301*

***Abstract*— The best entertainment for the younger generation now is given in the form of Social Networking sites. The Online Social Networks (OSN) mainly helps an individual to connect with their friends, family and the society online in order to gather and share new experiences with others. Now-a-days, the OSNs are facing the problem of the people posting the indecent messages on any individual's wall which annoys other people on seeing them. In order to filter those unbearable messages a system called Machine Learning is introduced. The aim of the present work is therefore to propose and experimentally evaluate an automated system, called Filtered Wall (FW), able to filter unwanted messages from OSN user walls. We exploit Machine learning (ML) text categorization techniques to automatically assign with each short text messages a set of categories based on its content. The major efforts in building a robust short text classifier (STC) are concentrated in the extraction and selection of a set characterizing and discriminating features.**

***Keywords*— Online social networks, information filtering, short text classification, and policy-based personalization.***

## I. INTRODUCTION

Online Social Networks (OSNs) is mainly used as an interactive medium to communicate, share, a considerable amount of human life information. OSN is mainly used to share several types of content, including text, image, audio, and video data. Online Social Network is a platform to build social networks (or) social relations among people who, for example, share interest, picture, text and real time connections. A social network service consists of each user having his own profile, his social links, and variety of additional services. It is web based service that allows individuals to create a public profile, to create a list of users with whom to share connection and to view the connection within the system. Some of the social networks which are mainly used to connect with friends are: Face book, Google+, YouTube, Twitter widely used worldwide.

Web content Mining is used to discover useful and relevant information from a large amount of Data. In OSN's, information filtering can be used for a different purpose. This is due to fact that in OSN's there is the possibility of posting (or) commenting other posts on particular public (or) private areas called *Walls*. Information filtering is mainly used to give user the ability to control the message written on their own walls by filtering out unwanted messages.

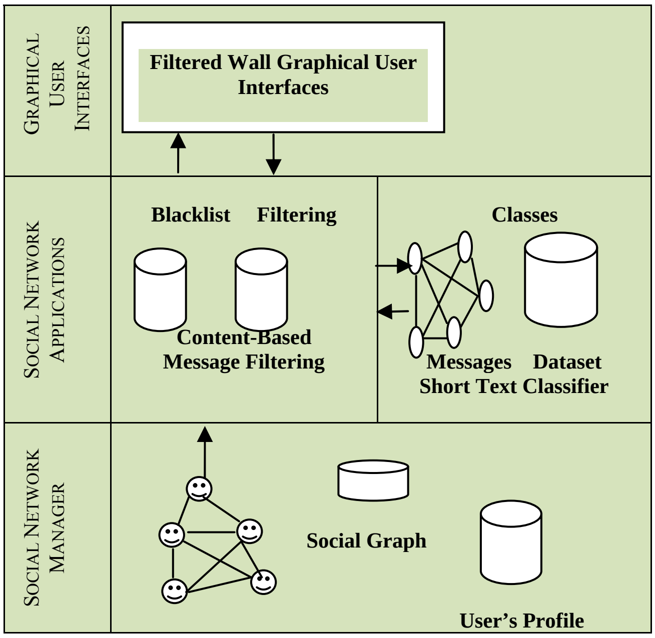

GRAPHICAL USER INTERFACES
Filtered Wall Graphical User Interfaces
SOCIAL NETWORK APPLICATIONS
Blacklist
Filtering
Content-Based Message Filtering
Classes
Messages
Dataset
Short Text Classifier
SOCIAL NETWORK MANAGER
Social Graph
User's Profile

Fig 1: OSN General Architecture

Today OSN's provide little support to prevent unwanted messages on user walls. For example face book allows users to state who is allowed to insert messages in their walls (i.e) friends, friends of friends, defined group of friends. Filtered wall is used to filter unwanted messages from OSN user walls. We used Machine Learning text categorization technique to automatically categorize each short text messages based on its content. We base the overall short classification strategy on Radial Basis Function Networks (RBFN) for their proven capabilities in acting as soft classifiers in managing noisy data
and intrinsically vague classes. We use the neural model RBFN categorizes as Neural and Non-neural FR filtering rules by which it can state what contents should not be displayed on their walls. In addition, the system provides the user defined Blacklists that is mainly used to temporarily prevent to post any kind of message on a user wall.

**Social Network Manager:** To provide the basic OSN functionalities (i.e) Profile and relationship management.

**Social Network Applications:** To provide external social network applications.

**Graphical User Interfaces:** User interacts with the system.

**Content Based Filtering:** It is mainly used to select information item based on the correlation between the content of the items and the user preferences.

**Filtering:** It is mainly used to filter the unwanted messages using Blacklists.

## II. RELATED WORK

M. Chau and H. Chen [2] describes as the Web continues to grow, it has become increasingly difficult to search for relevant information using traditional search engines. Topic-specific search engines provide an alternative way to support efficient information retrieval on the Web by providing more precise and customized searching in various domains. However, developers of topic-specific search engines need to address two issues: how to locate relevant documents (URLs) on the Web and how to filter out irrelevant documents from a set of documents collected from the Web. This paper reports our research in addressing the second issue. We propose a machine-learning-based approach that combines Web content analysis and Web structure analysis. We represent each Web page by a set of content-based and link-based features, which can be used as the input for various machine learning algorithms. The proposed approach was implemented using both a feed forward/back propagation neural network and a support vector machine. Two experiments were designed and conducted to compare the proposed Web-feature approach with two existing Web page filtering methods - a keyword-based approach and a lexicon-based approach. The experimental results showed that the proposed approach in general performed better than the benchmark approaches, especially when the number of training documents was small. The proposed approaches can be applied in topic-specific search engine development and other Web applications such as Web content management.

R.J. Mooney and L. Roy describe [3] Recommender systems improve access to relevant products and information by making personalized suggestions based on previous examples of a user's likes and dislikes. Most existing recommender systems use social filtering methods that base recommendations on other users' preferences. By contrast, content-based methods use information about an item itself to make suggestions. This approach has the advantage of being able to recommend previously unrated items to users with unique interests and to provide explanations for its recommendations. We describe a content-based book recommending system that utilizes information extraction and a machine-learning algorithm for text categorization. Initial experimental results demonstrate that this approach can produce accurate recommendations. These experiments are based on ratings from random samplings of items and we discuss problems with previous experiments that employ skewed samples of user-selected examples to evaluate performance.

F. Sebastiani describes The automated categorization[4] (or classification) of texts into predefined categories has witnessed a booming interest in the last ten years, due to the increased availability of documents in digital form and the ensuing need to organize them. In the research community the dominant approach to this problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of pre-classified documents, the characteristics of the categories.

The advantages of this approach over the knowledge engineering approach (consisting in the manual definition of a classifier by domain experts) are a very good effectiveness, considerable savings in terms of expert labor power, and straightforward portability to different domains. This survey discusses the main approaches to text categorization that fall within the machine learning paradigm. We will discuss in detail issues pertaining to three different problems, namely document representation, classifier construction, and classifier evaluation.

M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari [5] this paper proposes a system enforcing content-based message filtering for On-line Social Networks (OSNs). The system allows OSN users to have a direct control on the messages posted on their walls. This is achieved through a flexible rule-based system, that allows a user to customize the filtering criteria to be applied to their walls, and a Machine Learning based soft classifier automatically labeling messages in support of content-based filtering.

## III. EXINSTING SYSTEM

Indeed, Today OSNs provide very little support to prevent unwanted messages on user walls. For example, Face book allows users to state who is allowed to insert messages in their walls (i.e., friends, friends of friends, or defined groups of friends). However, no content- based preferences are supported and therefore it is not possible to prevent undesired messages, such as political or vulgar ones, no matter who posts them.

### *A. Disadvantages of Existing System*

1. Even though the Social Networks today, have the restrictions on the users who can post and comment on any user's wall, they do not have any restrictions on what they post. So, some people will use the indecent and vulgar words in commenting on the public posts.
2. Providing this service is not only a matter of using previously defined web content mining techniques for a different application, rather it requires to design ad hoc classification strategies.

## IV. NATURE OF WORK

Machine learning (ML) is used as text categorization techniques to automatically assign each short text message with in a set of categories based on its content. The major efforts in building a robust Short Text Classifier (STC) concentrate in the extraction and selection of a set characterizing and discriminating features. Here, a database of the categorized words is built and it is used to check the words if it has any indecent words. If the message consists of any vulgar words, then they will be sent to the Blacklists to filter out those words from the message. Finally, the message without the indecent words will be posted in the user's wall on the result of the content-based-filtering technique.

### A. *Advantages of proposed work*

1. A system automatically filters unwanted messages using the blacklists on the basis of both message content and the message creator relationships and characteristics.
2. Major difference include , a different semantics for filtering rules to better fit the considered domain, to help the users Filtering Rules(FRs) specification, the extension of the set of features considered in the classification process.

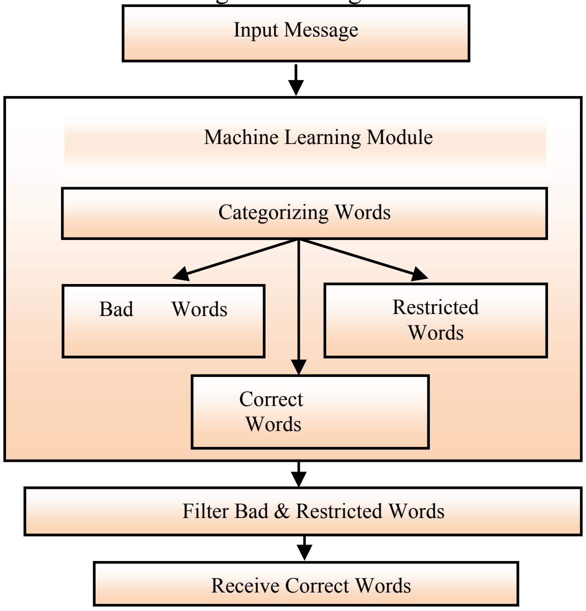
## V. Methodology

The users will create and manage their own "groups" (like the new Face book groups' pages). Each group has a homepage that provides a place for subscribers to post and share (by posting messages, images, etc.) and a block that provides basic information about the group. Users can also enable additional features in their owned page like view friends list and add friends by using friend's requests as well as share their images with selected group's members. The status of their friends' requests should also be updated here.

### A. *Filtering process*

In defining the language for FRs specification, we consider three main issues that, in our opinion, affect a message filtering decision. First, in OSNs like in everyday life, the same message may have different meanings and relevance based on who writes it. As a consequence, FRs should allow users to state constraints on message creators. Creators on which a FR applies can be selected on the basis of several different criteria; one of the most relevant is by imposing conditions on their profile's attributes. In such a way it is, for instance, possible to define rules applying only to young creators or to creators with a given religious/political view. Given the social network scenario, creators may also be identified by exploiting information on their social graph. This implies to state conditions on type, depth and trust values of the relationship(s) creators should be involved in order to apply them the specified rules. Fig.2. shows the filtering process.

Fig 2 : Filtering Process



The problem of setting thresholds to filter rules is also addressed, by conceiving and implementing within FW, an Online Setup Assistant (OSA) procedure. For each message, the user tells the system, the decision to accept or reject the message. The collection and processing of user decisions on an adequate set of messages distributed over all the classes allows computing customized thresholds representing the user attitude in accepting or rejecting certain contents. Such messages are selected according to the following process. A certain amount of non neutral messages taken from a fraction of the dataset and not belonging to the training/test sets, are classified by the ML in order to have, for each message, the second level class membership values.

### B. *Blacklisting Process*

A further component of our system is a Blacklist (BL) mechanism to avoid messages from undesired creators, independent from their contents. BL is directly managed by the system, which should be able to determine who are the users to be inserted in the BL and decide when user's retention in the BL is finished. To enhance flexibility, such information is given to the system through a set of rules, hereafter called BL rules. Such rules are not defined by the Social Network Management, therefore they are not meant as general high level directives to be applied to the whole community. Rather, we decide to let the users themselves, i.e., the wall's owners to specify BL rules regulating who has to be banned from their walls and for how long. Therefore, a user might be banned from a wall, and at the same time, he will not be able to post in the wall.

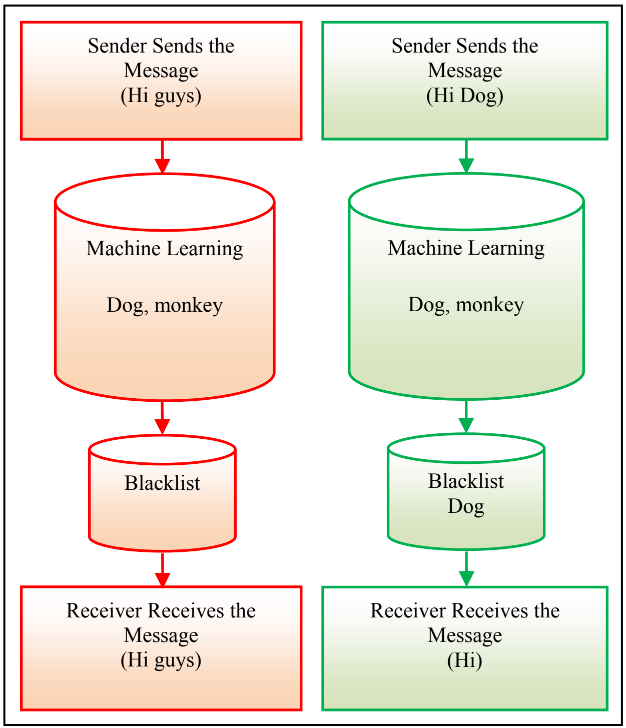

Sender Sends the Message (Hi guys)
Sender Sends the Message (Hi Dog)
Machine Learning
Dog, monkey
Machine Learning
Dog, monkey
Blacklist
Blacklist
Dog
Receiver Receives the Message (Hi guys)
Receiver Receives the Message (Hi)

Fig 3: Blacklist Process

Similar to FRs, our BL rules make the wall owner able to identify users to be blocked according to their profiles as well as their relationships in the OSN. Therefore, by means of a BL rule, wall owners are for example able to ban from their walls, users they do not directly know (i.e., with which they have only indirect relationships), or users that are friend of a given person as they may have a bad opinion of this person. This banning can be adopted for an undetermined time period or for a specific time window. Moreover, banning criteria may also take into account users' behavior in the OSN. More precisely, among possible information denoting users' bad behavior we have focused on two main measures. The first is related to the principle that if within a given time interval a user has been inserted into a BL for several times, say greater than a given threshold, he/she might deserve to stay in the BL for another while, as his/her behavior is not improved. This principle works for those users that have been already inserted in the considered BL at least one time. In contrast, to catch new bad behaviors, we use the Relative Frequency (RF) that let the system be able to detect those users whose messages continue to fail the Filtering Rules. The two measures can be computed either locally, that is, by considering only the messages and/or the BL of the user specifying the BL rule or globally, that is, by considering all OSN users walls and/or BLs.

The admin manages all the user's information including posting comments in the user status box. Each unwanted message has an alert from admin that provides a place for post and share for the respective user walls. And admin can see blocked message from the users and also that provides information about the user who used the blocked message. Admin can also enable additional features in their owned page like user list, adding unwanted message, update unwanted messages, Blocked users list and finally filter performance graph. And also in this module, we show the performance evaluation of the system in the graph.

### C. Algorithm Used

```
Step 1    Start
Step 2    A User tries post the message in a wall.
Step 3    Machine learning checks each word of the
          message.
Step 4    If (Words = = Good Words)
Step 5    Message is posted on the wall.
Step 6    Else if(Words = = Bad Words)
Step 7    Reject Bad Words using Blacklist and post
          the filtered message on the wall.
Step 8    Stop
```

As said earlier, the Machine Learning is a system which can learn from the data and take decisions based on the learned data. For example, a Machine Learning System in the Email Inbox can be used to learn and distinguish the emails received in the inbox between spam or non-spam emails. Similarly the Machine Learning here traces the posted messages for the good and the illegal words used in the wall by the public users.

The above algorithm represents the concept of Machine Learning with the Blacklist. Firstly, a user is showing his interest in posting or commenting in other person's wall regardless of their relationship. He can post any message there without the filtering technique. But the Machine Learning here learns the message which is yet to be posted and finds whether it contains any vulgar or illegal words in it. If it can't find any illegal or vulgar words, then the system allows the message to be posted on the wall. If it finds any illegal or vulgar words in that message while learning it, then it will remove the vulgar words from the message and then insert those words in the Blacklist which stores the indecent words in it. Finally the system prints the message without the indecent words. This mechanism helps in preventing the users to get annoyed by the vulgar words in a public wall of the Social Networking Sites. It does not prevent the unknown users from posting their messages; rather, it helps in preventing the obscenity with the vulgar words.

## VI. Experimental Evaluation

An experiment is also conducted with the Machine Learning Technique with the use of good and bad words in the messages posted on the Social Networking Site Wall. It is conducted with the consideration of the authorized and unauthorized persons taking part in the posts and comments. The graph shows that both the authorized and unauthorized persons use any kind of words in posting the messages.

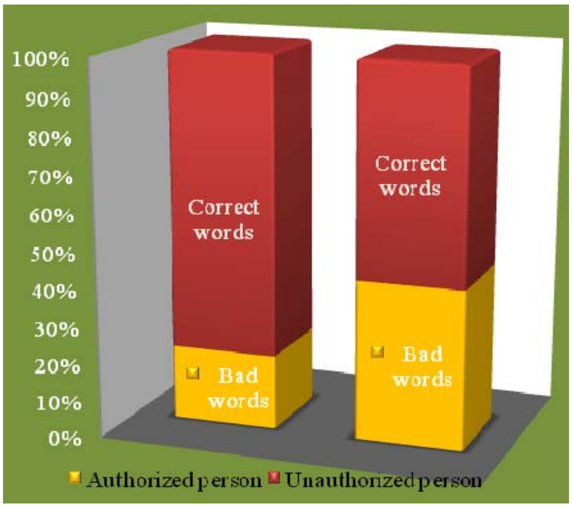


Fig 4: Graphical Representation

$$\text{Good Words} = \sum_{i=1}^{n} W_i \{M\} \neq \sum_{w=1}^{n} BL_w$$

$$\text{Bad Words} = \sum_{i=1}^{n} W_i \{M\} = \sum_{w=1}^{n} BL_w$$

Where, $W_i$ $i^{th}$ word of the message M
$M$ Input Message
$BL_w$ $w^{th}$ word in the Blacklist

For Example,
Blacklist words
$BL_w$ = {Dog, Monkey, Buffalo, Donkey}

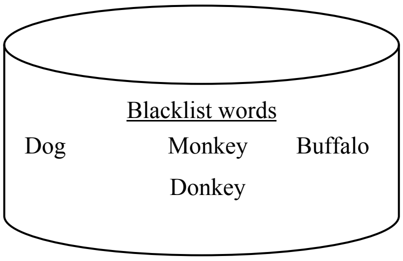


Fig 5: Blacklist Words

No. of Input Messages

$$M = \{Dog, Monkey, Kutty, Baby\}$$

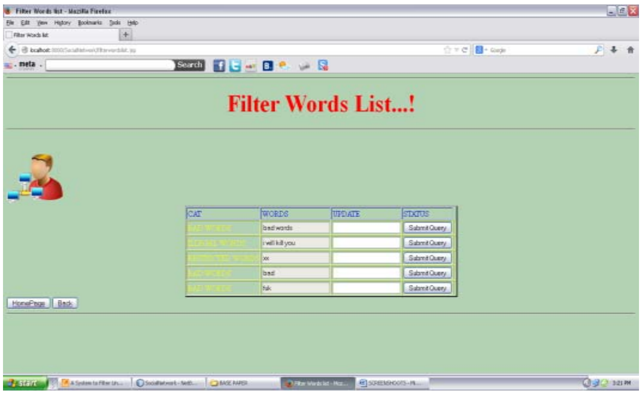
TABLE I. INPUT WORDS

| Input words | |
|---|---|
| $M_1$ | Hi Dog |
| $M_2$ | Monkey |
| $M_3$ | Buffalo |
| $M_4$ | Hi da Donkey what doing |

$$OP \Rightarrow (BL == M_1) = Hi$$
$$OP \Rightarrow (BL == M_2) = No\ Message\ Received$$
$$OP \Rightarrow (BL == M_3) = No\ Message\ Received$$
$$OP \Rightarrow (BL == M_4) = Hi\ da\ what\ doing$$

Where, OP Output
$M_i$ Message Input

### A. *Results*

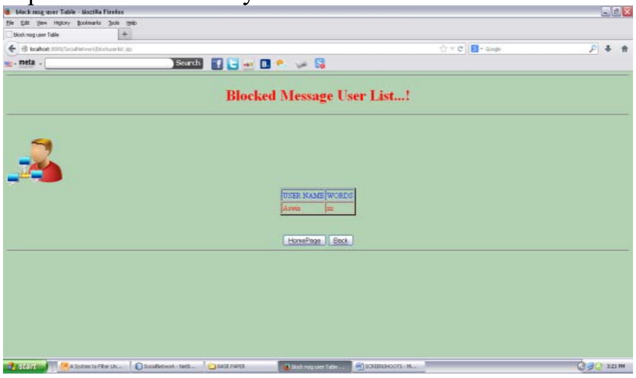


Filter Words List...!

| CAT | WORDS | UPDATE | STATUS |
|---|---|---|---|
| BAD WORDS | bad words | | Submit Query |
| [illegible] | i will kill you | | Submit Query |
| [illegible] | xx | | Submit Query |
| BAD WORDS | bad | | Submit Query |
| BAD WORDS | fuk | | Submit Query |

Fig 6: Filter word list

Fig. 6 shows the unwanted or bad words list stored by the system with the process of Blacklisting Process. This proves that not only the unauthorized persons can be prevented from posting of the messages in a person's wall. But the vulgar words can also be filtered from the message to prevent the obscenity.

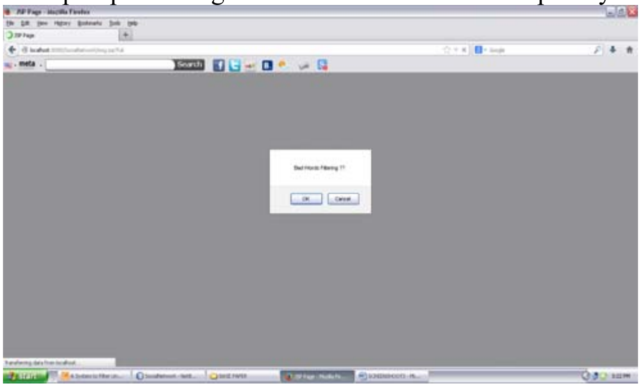


Blocked Message User List...!

| USER NAME | WORDS |
|---|---|
| Arwin | xx |

Fig 7: Blocked user list

Fig. 7 represents the Blocked User List. The Blocked User List consists of the Users who are not in the friends list of a user and who are not known to the person. These users can be filtered out and stored in a list for the system to help in preventing these users from further complexity.



Bad Words Filtering ??

Fig 8: Blocking unwanted words

Fig. 8 represents that the system will query to the user about blocking the unwanted words. If the user is not interested in showing these unwanted words, then he can reject those words from the message in showing up on his wall. If he wants the message to be printed as it is, then he can unblock the unwanted words in the messages. Then the posts will be as his wish.

## VII. CONCLUSION

In this paper, a system to prevent the indecent messages from the Social Networking site walls has been presented. The Usage of Machine Learning has given higher results to the system to trace the messages and the users to distinguish between the good and bad messages and the authorized and unauthorized users in the Social Networking User Profiles automatically. Thus the Machine Learning Technique plays a vital role in this paper in order to generate the blacklist of the bad words and the unauthorized users. The user has to update his privacy setting in his account in order to add this method to prevent the obscenity in his public profile. In this context, a statistical analysis has been conducted to provide the usage of the good and bad words by the persons in the sites. Overall, the obscenity of the users has been prevented.

## REFERENCES

[1] A. Adomavicius and G. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 6, pp. 734-749, June 2005.

[2] M. Chau and H. Chen, "A Machine Learning Approach to Web Page Filtering Using Content and Structure Analysis," Decision Support Systems, vol. 44, no. 2, pp. 482-494, 2008.

[3] R.J. Mooney and L. Roy, "Content-Based Book Recommending Using Learning for Text Categorization," Proc. Fifth ACM Conf. Digital Libraries, pp. 195-204, 2000.

[4] F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

[5] M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari, "Content-Based Filtering in On-Line Social Networks," Proc. ECML/PKDD Workshop Privacy and Security Issues in Data Mining and Machine Learning (PSDML '10), 2010.

[6] N.J. Belkin and W.B. Croft, "Information Filtering and Information Retrieval: Two Sides of the Same Coin?" Comm. ACM, vol. 35, no. 12, pp. 29-38, 1992.

[7] P.J. Denning, “Electronic Junk,” Comm. ACM, vol. 25, no. 3, pp. 163-165, 1982.

[8] P.W. Foltz and S.T. Dumais, “Personalized Information Delivery: An Analysis of Information Filtering Methods,” Comm. ACM, vol. 35, no. 12, pp. 51-60, 1992.

[9] P.S. Jacobs and L.F. Rau, “Scisor: Extracting Information from On-Line News,” Comm. ACM, vol. 33, no. 11, pp. 88-97, 1990.

[10] S. Pollock, “A Rule-Based Message Filtering System,” ACM Trans. Office Information Systems, vol. 6, no. 3, pp. 232-254, 1988.

[11] P.E. Baclace, “Competitive Agents for Information Filtering,” Comm. ACM, vol. 35, no. 12, p. 50, 1992.

[12] P.J. Hayes, P.M. Andersen, I.B. Nirenburg, and L.M. Schmandt, “Tcs: A Shell for Content-Based Text Categorization,” Proc. Sixth IEEE Conf. Artificial Intelligence Applications (CAIA ’90), pp. 320-326, 1990.

[13] G. Amati and F. Crestani, “Probabilistic Learning for Selective Dissemination of Information,” Information Processing and Management, vol. 35, no. 5, pp. 633- 54, 1999.

[14] M.J. Pazzani and D. Billsus, “Learning and Revising User Profiles: The Identification of Interesting Web Sites,” Machine Learning, vol. 27, no. 3, pp. 313-331, 1997.

[15] Y. Zhang and J. Callan, “Maximum Likelihood Estimation for Filtering Thresholds,” Proc. 24th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 294-302, 2001.

[16] C. Apte, F. Damerau, S.M. Weiss, D. Sholom, and M. Weiss, “Automated Learning of Decision Rules for Text Categorization,” Trans. Information Systems, vol. 12, no. 3, pp. 233-251, 1994.

[17] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, “Inductive Learning Algorithms and Representations for Text Categorization,” Proc. Seventh Int’l Conf. Information and Knowledge Management (CIKM ’98), pp. 148-155, 1998.

[18] D.D. Lewis, “An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task,” Proc. 15th ACM Int’l Conf. Research and Development in Information Retrieval (SIGIR ’92), N.J. Belkin, P. Ingwersen, and A.M. Pejtersen, eds., pp. 37-50, 1992.

[19] R.E. Schapire and Y. Singer, “Boostexter: A Boosting-Based System for Text Categorization,” Machine Learning, vol. 39, nos. 2/3, pp. 135-168, 2000.

[20] H. Schu¨ tze, D.A. Hull, and J.O. Pedersen, “A Comparison of Classifiers and Document Representations for the Routing Problem,” Proc. 18th Ann. ACM/SIGIR Conf. Research and Development in Information Retrieval , pp. 229-237, 1995.

**Babu** is an Assistant Professor in the Dept. of Computer Science and Engineering at M.R.K.Institute of Technology, Kattumannarkoil, India from 2013. He received his B.Tech degree at E.S.College of Engineering and Technology from Anna University in 2011, Chennai and the M.Tech degree at Anna University Regional Centre, Coimbatore from Anna University, and Chennai in 2013.

**Charles** is an Assistant Professor in the Dept. of Computer Science and Engineering at M.R.K.Institute of Technology, Kattumannarkoil, India from 2013. He received his B.Tech degree at Annai Teresa College of Engineering and Technology from Anna University in 2009, Chennai and the M.Tech degree at Prist University Pondicherry in 2012.