Rule Criteria Selection Based System To Filter Unwanted Messages And Images From OSN Users Private Space

Miss. Swapnali S. Gawali[1], Prof. D. B. Kshirsagar[2]
PG Student, Department of Computer Engg, SRES Sanjivani COE, Kopargaon, Savitribai Phule Pune University, Pune. India[1]. gswap13@gmail.com
Professor D. B. Kshirsagar, SRES Sanjivani COE, Kopargaon, Savitribai Phule Pune University, Pune. India[2].hodcompcoe@sanjivani.org.in

Abstract- *The Online Social Networking (OSN) is the need of today's world. Most of people wants to keep update their profiles in terms to share their thoughts through the OSN in the form of text, images, audio, video etc. As many people share their data on OSN the data size also increasing proportionally. With the increasing the people's share in the OSN in terms of their data so that the data also facing the problems like to protection of registered user's own wall from vulgar messages and bad images which will publish by other users. As per the standard Online Social Networking websites the user does not have direct control for their own wall. The sharing and display of irrelevant text or images should be a part of privacy control. To keep this issues in mind we are implementing the solution with more privacy filtering techniques using support vector machine and skin detection. Also design OCR to detect word from images & filter them.*

Keywords – Online Social Network, Filter Wall, Content Based Message Filtering (CBMF), Filtering Rule (FR), BlackList (BL), Machine Learning, Radial Basis Function Network (RBFN).

I. INTRODUCTION

Online Social Networks are more famous and interactive medium to connect many peoples, sharing thoughts, communicating and discuss events. They exchange many content in different form like text, images, audio, video data. In online social network the meaning of a wall is user's space where the possibility of posting and commenting of other post particular public or private. Information filtering in OSNs gives the ability to users to automatically control the messages and images written on their walls, by filtering unwanted messages and images. Indeed, recent OSNs prevent unwanted messages in very little manner. Facebook allows user to state who is allowed to insert messages in their walls(i.e Friends, Friends of friend, defined groups). There in no content based preferences supported, thus making it impossible to prevent undesired messages, no matter who post them. From a security point of view, social networks have unique characteristics. First, information access and interaction is based on trust. Users typically share a substantial amount of personal information with their friends. This information may be public or not. If it is not public, access to it is regulated by a network of trust. In this case, a user allows only friends to view the information regarding herself. Moreover, it often happens that users, to gain popularity, accept any friendship request they receive, exposing their personal information to unknown people.

We are designing machine learning text categorization techniques to categorized each short text message on its content. Design short text classifier which focuses on extraction and selection of a set of characterizing and recognize features. Derived externally knowledge related to the context from which message originate is add to the short texts along with original set of features, derived from derived internally properties. We have designed the system in stages. In first stage, we categories the messages by using support vector machine. We have chosen this because it has high dimensional input space. When learning text classifiers user has many (more than 10000) features. Since SVMs use over fitting protection, which does not necessarily depend on the number of features, they have the potential to handle these large feature spaces. Few irrelevant features, this is other way to avoid these high dimensional input spaces is to assume that most of the features are irrelevant [8]. It categorizes messages as Neutral and Non neutral and in second stage non neutral messages are classified depending upon category. The system including classification techniques provides rule layer to specify Filtering rules in flexible language. This filtering rule helps users to state, which content should be unwanted to display on their walls. According to users requirements filtering criteria's combined and customized. The proposed system provides not only classification and filtering rule but also provide BlackList (BL). BL is list of users that are temporarily prevented to post any kind of messages on user wall. The second part of the designed system is to filter unwanted images from user's wall. We can achieve this filtering by using skin detecting method which detects bad human images.

Also consider the images which consist malicious text to detect that text we design OCR technique. And after detection we filter that text. The rest of the paper is organized as follows. Section II presents the literature survey of the existing filtering techniques, Section III describes system architecture, Section IV Implementation of the system. Section V describes the Result of the implemented system, while Section VI concludes the paper.

II. LITERATURE SURVEY

The content based filtering is an emergent area of research. Many researchers has working on this area.

Macro Vanetti *et. al.* [1] proposed Filtered Wall architecture for OSN where users have been allowed to publish only filtered content. They provided system where users able to controlled their private space. Machine learning based soft classifier which automatically labelled the messages in support content based filtering. Author has only focused on text content.

Nicholas J. Belkin and W. Bruce Croft [2] has proposed Information filtering systems were designed to classify a stream of dynamically generated information dispatched asynchronously by an information producer and presented to the user those information that were likely to satisfy their requirements. In content-based filtering each user has assumed to operate independently. As in result, a content-based filtering system selected information items based on the correlation between the content of the items and the user preferences as opposed to a collaborative filtering system that has been chosen items based on the correlation between people with similar preferences.

Zelikovitz and Hirsh [3] attempted to improve the classification of short text strings by developed a semi supervised learning strategy based on a combination of labelled training data plus a secondary corpus of unlabeled but related longer documents. This solution is inapplicable in our domain in which short messages are not summary or part of longer semantically related documents.

A different approach has been proposed by Bobicev and Sokolova [4] that circumvent the problem of error-prone feature construction by adopting a statistical learning method that were performed reasonably well without feature engineering. However, this method, named Prediction by partial Mapping, produced a language model that used in probabilistic text classifiers which are hard classifiers in nature and do not easily integrate soft, multi-membership paradigm.

B. Sriram *et.al.* [5] has proposed a classification method to categorize short text messages in order to avoid overwhelming users of micro blogging services by raw data.

Golbeck and Kuter [6] proposed an application, called FilmTrust, that exploited OSN trust relationships and provenance information to personalize access to the website.

Christian Platzer *et.al.*[7] proposed skin sheriff, which was trainable tools which helped to automatically detect with high precision and recall a pornographic part in images. To rate unknown arbitrary images by combining skin sheriff novel skin detection mechanism with a support vector machine which is highly dynamic in nature. Trained detection engine used to target images on specific domain.

Claudiu et al.[9] has consider committee-based classifiers of isolated handwritten characters are the first on par with human performance and can be used as basic building blocks of any OCR system (all our results were achieved by software running on powerful yet cheap gaming cards).

Georgios et al.[10] has presented a methodology for off-line hand written character recognition. The proposed methodology relies on a new feature extraction technique based on recursive subdivisions of the character image so that the resulting sub-images at each iteration have balanced (approximately equal) numbers of foreground pixels, as far as this is possible. Feature extraction is followed by a two-stage classification scheme based on the level of granularity of the feature extraction method. Classes with high values in the confusion matrix are merged at a certain level and for each group of merged classes, granularity features from the level that best distinguishes them are employed.

III. SYSTEM ARCHITECTURE

To overcome the limitations of the existing system we implementing system having an architecture shown in figure1.

The three tier architecture in support of OSN services (Figure 1).

Layer 1 Social Network Manager (SNM): The profile and relationship management is main task of Social network management layer. It contain the information of users profiles and provides this information to the second layer for applying filtering rules (FR) and blacklists (BL).

Layer 2 Social Network Application (SNA): These second layers apply for filtering purpose. This layer consist Content Base Message Filtering (CBMF) and a short text classifier is most important layer. The classifier classifies each message according to its content and CBMF filters the messages according to filtering rule and blacklist given by the user.

Layer 3 Graphical User Interface (GUI): Third layer graphical user interface where user enter his input and wait to see published wall messages.

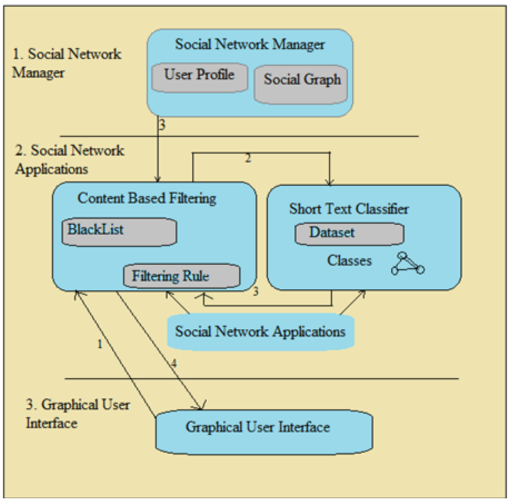

Figure 1. System Architecture.

Following path has been followed by an architecture to filter the messages and publish content

- User tries to post content on his own or friends' wall, which is intercepted by FW.
- Extraction of meta-data from content of message by using Machine-Learning based Classifier.
- Meta data provided by classifier and information provided by SNM layer use to enforce the filtering rule and BL.
- Depending on result of third steps content able to published or not able to publish decides.

IV. IMPLEMENTATION

Detailed work break down structure of our system has been shown in figure. 2.

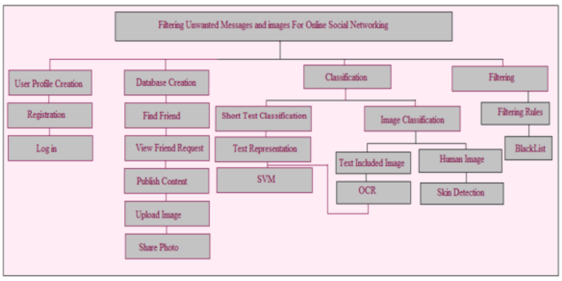

Figure 2. System Breakdown Structure

Module 1: User Profile Creation

In this module, user can register their details like name, Password, gender, age, and then only registered users can login to system.

Module 2: Database Creation

In this module database is created for the user to handle the user account.

- Find Friend - User can find for friends and can view their details.
- View Request - User can see the Friend request and accept or reject the request.
- Upload Image - User can upload the image.
- Share photo - User share the photo.
- Publish content - Use want to publish the content they want to publish.

Module 3 Classification

Information filtering systems are designed to classify a stream of dynamically generated message dispatched asynchronously by an message producer and present to the user those content that are likely to satisfy his/her requirements.

- *Short Text Classification*: Most of the established classification techniques used for text classification work well on datasets with large documents, but suffer when the documents in the corpus are short. To overcome this drawback design short text classification which focuses on extraction and selection of a set of characterizing and recognize features of short text. This classification used in hierarchical strategy. The first level will be classified with neutral and non neutral labels. The second level consider non neutral sentence for further processing.
- *Text Representation*: Classification performance is depends on the text representation of a document. This text representation is critical tasks strongly affect classification strategy. There are many features for text representation but we consider only three types of features. Document properties (DP), contextual features (CF) and BOW. BOW and DP used in content based filtering; endogenous that is, they consider information in text of messages to derived the representation of text. The source of information takes from outside the message but it directly or indirectly related to message itself is nothing but the exogenous knowledge. Understanding the semantics of message CF modelling introduced. DP features are considered known words and statistical properties. DP features heuristically consider, some domain specific criteria evaluate trial and error procedures are needed for some cases.

  1. Correct words: In this express the correct word to represent.
  2. Bad words: It express same as correct words but only "dirty words" will be determine.
  3. Stop Words: It collect the stop words from messages.
  4. Total words: It calculate the amount of words in message.
- *Support Vector Machine*: SVM [8] is a supervised machine learning algorithm. This algorithm can be used for both classification and regression challenge. Mostly, it is used in classification problems. In this, we plot each words as a point in n-dimensional space (where n is number of different words we have) with the value of each word being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two malicious words categories very well as in figure 3.

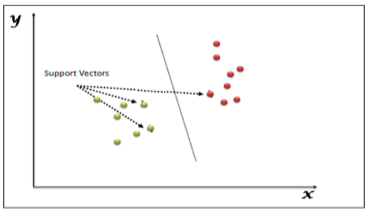

Figure 3 : Support Vector Machine

- *Image Classification :* Image classification detect the bad images to publish and not allow them to publish. This image classification considers the two types of images first one is human image and second one is image contains the text. Filtering human images we used skin detecting algorithm. Detecting and reorganization text on images used OCR algorithm.
- *Skin Detecting:*

  The aim of the Skin Detecting Component is to fetch all skin areas from an image. To determine bad images first classify and label all pixels separately and then mark them in a binary image. Classified skin area pixel label in grey and non skin pixels in white. The label skin and non skin area pixel is called skin map [7].

  Algorithm : Skin Detection

  Input : An image

  Output: filtered images(ignored if pornography)

```
function SkinDetect(img,imgwidth, imgheight)
Scale(img, width < 1000px)
AutoContrast(img)
skinmap ← NewImage(imgwidth, imgheight, white)
for all pixel in img do
R, G, B ← pixel
H, S, V← ConvertRGBtoHSV(R; G;B)
if IsSkin(R,G,B,H, S, V ) then
skinmap[pixelx; pixely]←grey
else
```

```
skinmap[pixelx; pixely] ←white
end if
end for
grey closing(skinmap; size (6; 6))
return skinmap
end function
```

- OCR

  Optical character recognition (OCR) is a process of converting a printed document or scanned page into ASCII characters that a computer can recognize. An algorithm for implementation of Optical Character Recognition (OCR) to translate images of typewritten or handwritten characters into electronically editable format by preserving font properties. OCR can do this by applying pattern matching algorithm. The recognized characters are stored in editable format. Thus OCR make the computer read the printed documents discarding noise. Following processes follow to implement OCR.

  1. Greyscale
  2. Feature Extraction
  3. Recognization of Pattern
  4. Recognization of Output

Module 4: Filter System

In filtering module system consider the filtering rule and BlackList. Filtering rule and blacklist both the set by user, as per that selection filtering process perform. Because of that users get own control on their OSN wall. This filtering process consist five types of categories like hate, vulgar, offensive, violence, sexual. All this categories included the different words.

- Filtering Rule: Filtering rule is selection of categories by user. This categories word user want to filter during the filtering process. Users set this rule any time. This is easiest to set filtering criteria. By setting rules by user system perform the filtering.
- BlackList: BlackList consist the block user information and User added filtering string. Filtering string is string added by user to filter that string for user wall messages. Users have own rights to block/unblock the person.

VI. RESULT

The performance of our system tested on manually added messages. This messages classified according to five categories then filtering of messages have been depends upon the users specified category they mention on profile. For each classification there are 30 words available in each category. It means there are total 150 words use to train the classifier. These train words cover categories including hate, vulgar, offensive, violence and sexual etc. Dataset two uses the images for filtering. This is a separate dataset as like word dataset. The dataset has contains 50 good images for training while other 50 images have malicious images. Dataset three uses the images included text has available. The dataset has contains near about 30 images for training. As soon as the malicious text into images has identified by optical character recognitions for filtering.

The performance of system is measure by Precision and recall. Precision is the proportion of returned messages that are targets, while recall is the proportion of target documents returned.

Precision = TP/TP+FP

Recall = TP/TP+FN

Where,
TP = Correct input correctly identify.
TN = Correct input incorrectly identify.
FP = Incorrect input correctly identify.
FN = Incorrect input incorrectly identify.

The following table 1 shows the result analysis of text messages filtering. In the performance analysis we had taken 300 good messages and 550 malicious messages. Figure 4. Shows the graph for Precision, Recall measured for text classification.

TABLE 1
Result For The Text Filtering System In Terms OF PRECISION (P), RECALL (R) for each categories

Categories	TP	TN	FP	FN	Precision	Recall
Good	240	30	40	20	0.85	0.92
Hate	70	10	20	10	0.77	0.87
Sexual	79	7	18	9	0.81	0.89
Offensive	79	7	13	11	0.85	0.87
Vulgar	81	8	8	13	0.91	0.86
Violence	80	6	17	7	0.82	0.91

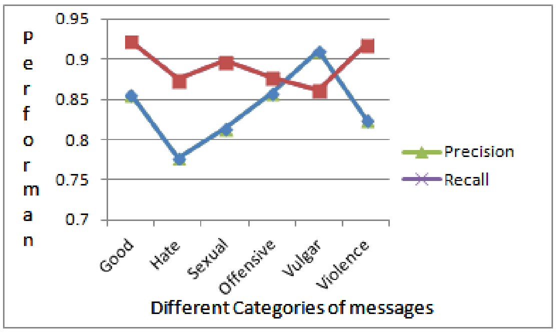

Figure 4 : Graph represent Precision, Recall for text classifition

In Images analysis we had taken 50 good images, 50 bad images and 10 text included images.

Table 2 shows the images analysis for good, bad & text included images. Figure 5. Shows Precision, Recall graph for images analysis.

TABLE 1
Result For The Text Filtering System In Terms OF PRECISION (P), RECALL (R) for each categories

Categories	TP	TN	FP	FN	Precision	Recall
Good Images	40	3	4	3	0.90	0.93
Bad Images	38	3	5	4	0.88	0.90
Text included Images	6	1	2	1	0.75	0.85

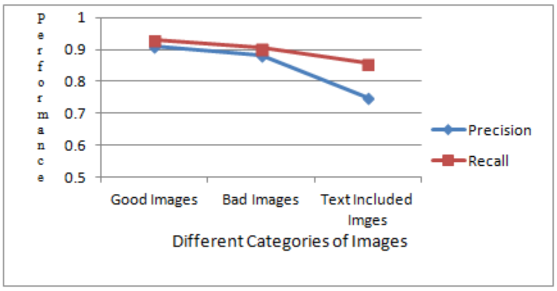

Figure 5: Graph For Images Analysis.

VII. CONCLUSION & FUTURE SCOPE

In this paper we are extending the method to provide unwanted message filtering for social networks. The system has filtered the messages by support vector machine while to filter the images system has utilizes the skin detecting algorithm. To make the system more favourable we have implemented optical character recognition in images. By implementing this various strategies in system we made the user much secured from the unwanted message as well image point of view. In the future work our system will consider the various ways of to generate the unwanted messages and filter them.

ACKNOWLEDGEMENT

I would like to take this opportunity to express my heartfelt thanks to my guide Prof. D.B. Kshirsagar for his esteemed guidance and encouragement, especially through difficult times. His suggestions broaden my vision and guided me to succeed in this work. I am also very grateful for his guidance and comments while designing part of my research paper and learnt many things under his leadership.

REFERENCES

[1] Marco Vanetti, Elisabetta Binaghi, Elena Ferrari, Barbara Carminati and Moreno Carullo, *"A System to Filter Unwanted Messages from OSN User Walls",* IEEE Transactions On Knowledge And Data Engineering Vol:25, pp. 1-15, 2013.

[2] N. J. Belkin and W. B. Croft,*"Information filtering and information retrieval: Two sides of the same coin?",* Communications of the ACM, vol. 35, no. 12, pp. 29-38, 1992.

[3] S. Zelikovitz and H. Hirsh, *"Improving short text classification using unlabeled background knowledge",* in Proceedings of 17th International Conference on Machine Learning (ICML-00), P. Langley, Ed. Stanford, US: Morgan Kaufmann Publishers, San Francisco, US, pp. 1183-1190, 2000.

[4] V. Bobicev and M. Sokolova, "*An effective and robust method for short text classification*", in AAAI, D. Fox and C. P. Gomes, Eds. AAAI Press, pp. 1444-1445, 2008.

[5] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "*Short text classification in twitter to improve information filtering*", in Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, pp. 841-842, 2010.

[6] J. Golbeck, "*Combining provenance with trust in social networks for semantic web content filtering, in Provenance and Annotation of Data, set.*", Lecture Notes in Computer Science, L. Moreau and I. Foster, Eds. Springer Berlin / Heidelberg, vol. 4145, pp. 101-108, 2006.

[7] Christian Platzer, Martin Stuetz, Martina Lindorfer, " *Skin Sheriff: A Machine Learning Solution for Detecting Explicit Images*", ACM 978-1-4503-280, June 2014.

[8]Thorsten Joachims, "*Text Categorization with Support Vector Machines: Learning with Many* Relevant Features", University at Dortmund Informatik LS8, Baroper Str. 301 44221 Dortmund, Germany.

[9]Dan ClaudiuCires¸an and Ueli Meier and Luca Maria Gambardella and Jurgen-Schmidhuber, "*Convolutional Neural Network Committees for Handwritten Character Classification,*" 2011 International Conference on Document Analysis and
Recognition, IEEE, 2011.

[10] GeorgiosVamvakas, Basilis Gatos, Stavros J. Perantonis, "*Handwritten character recognition through two-stage foreground sub-sampling,*" Pattern Recognition, Volume 43, Issue 8, August 2010.