‘CHATGAURD’-A system that ensures safe posting in social networking sites

Surya Prabha. K[1], Harini. N[2]

[1]Computer Science and Engineering,
Amrita School of Engineering, Coimbatore, India,
ksurya22m@gmail.com

[2]Computer Science and Engineering,
Amrita School of Engineering, Coimbatore, India,
n_harini@cb.amrita.edu

ABSTRACT - In this paper, we propose a system called ChatGaurd for safe posting of messages by allowing the users to have direct control on messages. This is accomplished through a scheme that ensures non-repudiation without compromise of privacy, and an integrated mechanism of filtering offensive words using machine learning approach. The system uses a list of blacklisted words in addition to a dataset of customized bad words using which analysis is carried out to categorize the message into positive, negative or neutral. The scheme has been illustrated on an application that is used by parents to effectively monitor the message posts of their wards on online networking sites (OSNs).

KEYWORDS - Online social networking sites, Message posts, Blacklist, Machine Learning, Mylist, Blind signature.

1. INTRODUCTION

Online social networking (OSN) is one of the most interactive medium to share, communicate and exchange information’s like text, audio, video, etc. Sites like Facebook, Orkut, Twitter and more helps to stay connected in people’s lives through status, photos and videos they post publicly. Social media is indispensible for all age groups with the convenience access provided by mobiles. 92% of teens use social media daily out of which 71% of teens uses more than one OSN. According to a study from Pew Research Center, more than half of teens are from age group 13-17. *Figure 1* shows the percentage of teens using services offered by social networking sites.

With the exponential growth in the number of novice people using internet, the hackers have found it a cake walk to take advantage of the good intentions of social activities. There are cases reported that say about users posting offensive or unwanted messages on other’s walls. This act may affect a victim psychologically, and the same is brought out in *Figure-2* which shows the comment of a celebrity about Chennai flood victims, which may psychologically affect Chennai residentswho have already become victims to flood.

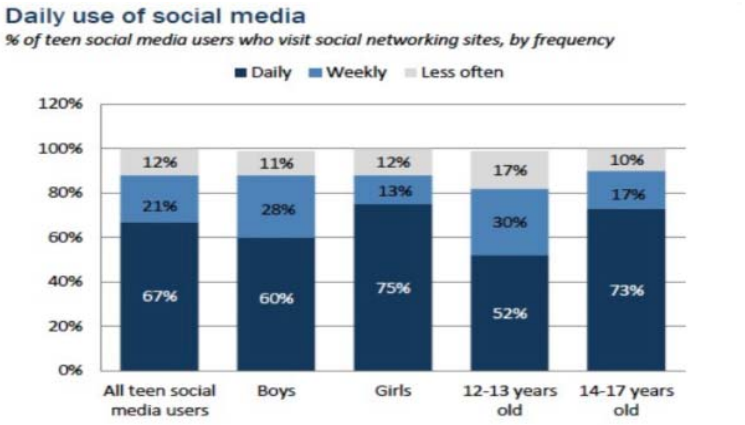

Figure 1- Percentage of teens who use OSN, by frequency

Figure 2- Celebrity from Mumbai post a racist thing on OSN, about Chennai flood victim.

Messaging service offered by social networking sites keeps most of the people connected. This could be for exchange ofmessages related to studies, research, interest, etc. Participants of social networking sites who are not aware of the deeds of hackers/crackers/cyber criminals may have to face problems related to their account management and conversations done using messaging services.

Research works [2, 7 and 14] provide methods for safe messaging using NLP (Natural Language Processing) tools. These works mainly filters the offensive contents in messages based on output provided by the NLP tool. Researches have demonstrated filtering of offensive post on shopping websites based on the comments provided about a particular product and its features. But, to the best of our knowledge we have not come across any work that concentrates on analyzing personal messages on individual's wall. This work focuses on providing a safe mode of messaging on personal communications for teenagers. The highlight of the work is that it does it without compromise in privacy. The contribution of this paper includes the following:

1. Quick comparison with the blacklisted words that facilitates tweeting without any delay.
2. Analysis of tweets using NLP tool to filter the offensive words that were not blacklisted.
3. Fast analysis using customized blacklisted words which each individual considers to be offensive (for example, a parent might not want their ward to use words such as stupid, lame, etc.. which had not looked up on as offensive words in broader sense).
4. Ensuring privacy and non-repudiation by using blind digital signature scheme.

2. RELATED WORKS

This work [13] by Sunil Yadav et al and S Das, D Rudrapal proposed a system which may allow OSN users to have a direct control on posting or commenting on their walls with the help of information filtering. This is achieved through text pattern matching system, that allows users to filter their open space and a privilege to add new words treated as unwanted.In [8] K Babu, P Charles proposed a system, which discusses on machine learning which categories the message based on its content and filtered wall to filter unwanted messages from users wall.[6] by Amruta kachole, S.D Jondhale presents a system which gives ability to users to control the messages posted on their own private space to avoid unwanted messages displayed by the use of Customizable Filtering Rules as well as machine learning approach and black list techniques are applied on user's wall.In [12] T.K Das, D.P Achariya and M.R Patra have explained the detailed work done in developing a system which can be used for the purpose of opinion analysis of a product or a service.In [9], Subhabrata Mukherjee presents a novel approach to identify features specific expressions of opinion in product reviews with different features and mixed emotions.[2] proposes a system which identifies posts which are aimed at hurting the sentiments of others and makes the user to rethink and hence refrain from posting the same using Stanford Core NLP tool and it also provides an effective algorithm that identifies and reduces the spam content in the user's message.

Most of the work discussed above focus on providing recommendation to users based on his/her likes (Recommendary system) such system collect individual preferences of users and use this information to receive post on their walls based on their like. Mostly such systems use an NLP analyzer to carry out the analysis.

2.1 FINDINGS

1. NLP tools can be used to perform analysis on messages.

2. To facilitate safe messaging, works use blacklist to filter offensive words.

3. Many system reported till date, drops off messages in a particular domain like customers feedback on products.

4. A work that performs customized filtering of offensive words on personal communication has not been proposed still.

3. PROPOSED SYSTEM

A detailed discussion of system is carried out in this section. This system operates at two logical layers namely Application layer and Security layer. Application layer focuses on detecting and dropping of offensive words using NLP tool and blacklist mechanism and the Security layer ensures non-repudiation.

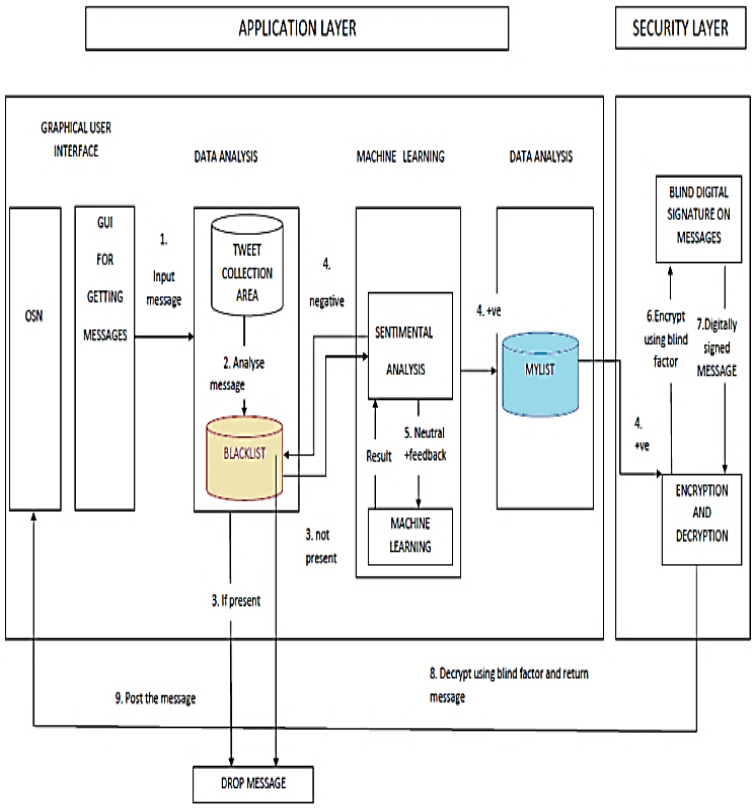

Figure 3- Chatgaurd architecture diagram.

3.1 FUNCTIONING OF APLLICATION LAYER

The usage of slang languages (offensive words), affects a person psychologically. To overcome the above mentioned problem such offensive words need to be filtered. The proposed system uses two lists for complete filtration of offensive words (Blacklist, Mylist).

1. Blacklist - Contains the huge list of blacklisted words that were commonly used by users and Google banned words and stored them in mongo DB .*Figure-4* Shows a sample set of blacklisted words used as dataset for experimentation.
2. Mylist – Social networking sites are one of the most interactive medium to communicate and share information's. Sometimes, the messages may be considered as offensive by individual's (which may not be offensive for others) and would want to remove it from the wall. So, a survey was conducted among a set of parents and a list was created which includes the words that the parents find it to be offensive and even they do not want others to post it in the walls of their wards. *Figure 6*- A, B, C show the report of the survey conducted. *Figure 5*- Represents the dataset that constitute mylist words which were derived from mining the survey report. *Figure 7*- show the working model of this system for the message containing Mylist word. *Figure 8*- show the working model of no mylist words.

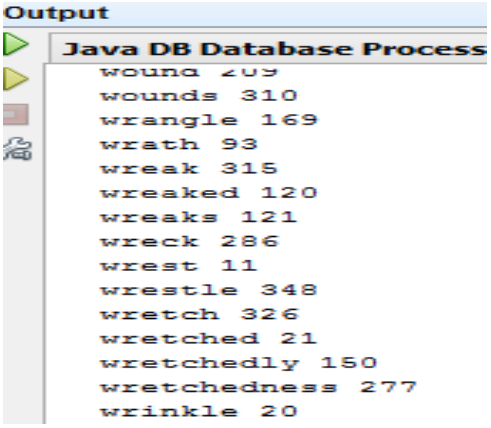

Figure 4 – blacklisted words in mongoDB

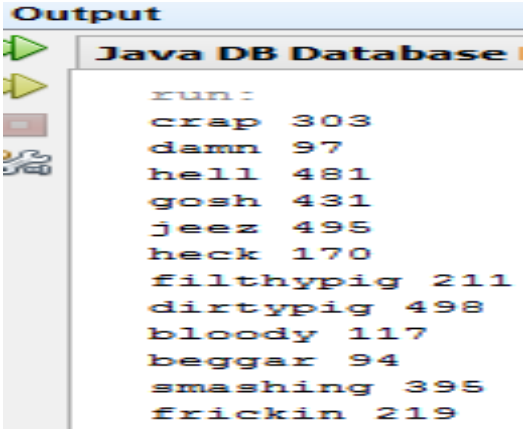

Figure 5- Mylist words

Figure 3- Shows the architecture of proposed method. The algorithm for proposed system as follows, Algorithm 1.0 specifies the steps carried out by 'ChatGaurd' system.

ALGORITHM 1.0 CHARTGAURD

INPUT – Message (m) to be checked. **OUTPUT**- Status of post. ***STEP 1*** – Start. ***STEP 2*** - Input m. ***STEP 3*** - Checks for availability of 'm' in Blacklist. ***STEP 4*** - Analysis by NLP tool If(m= positive) then go to step 5 else if(m= negative) then drop m else Get feedback from user and analyze. ***STEP 5*** – End if. ***STEP 6*** – End if. ***STEP 7*** - Checks for availability of 'm' in mylist ***STEP 8*** – Stop.

TABLE 1

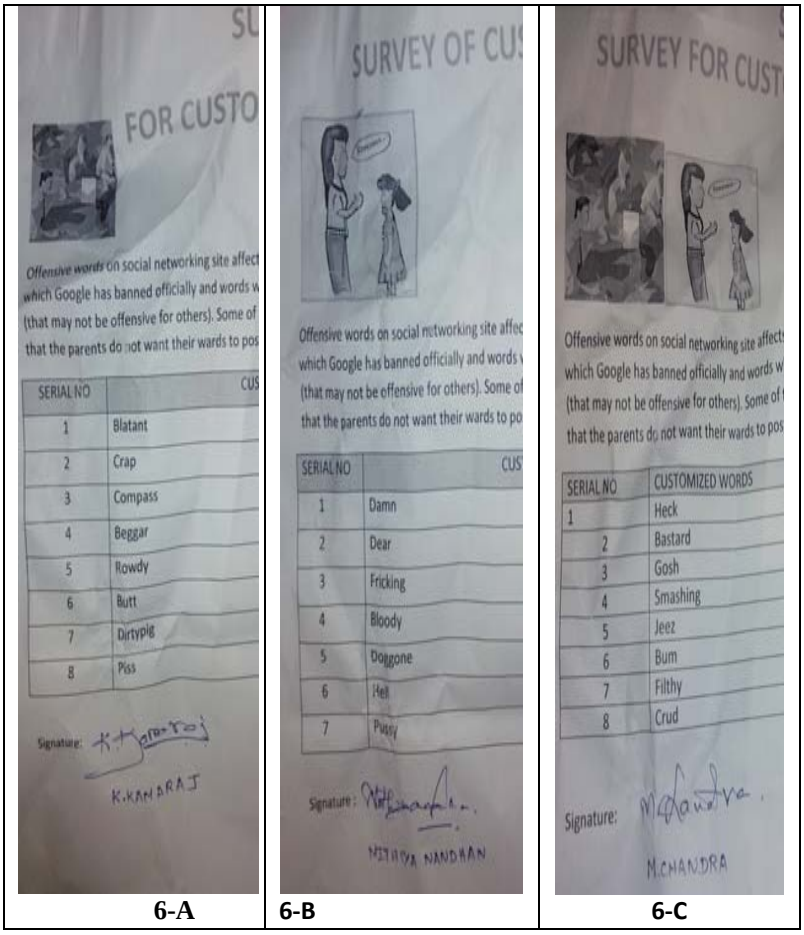

Figure 6-Survey Report

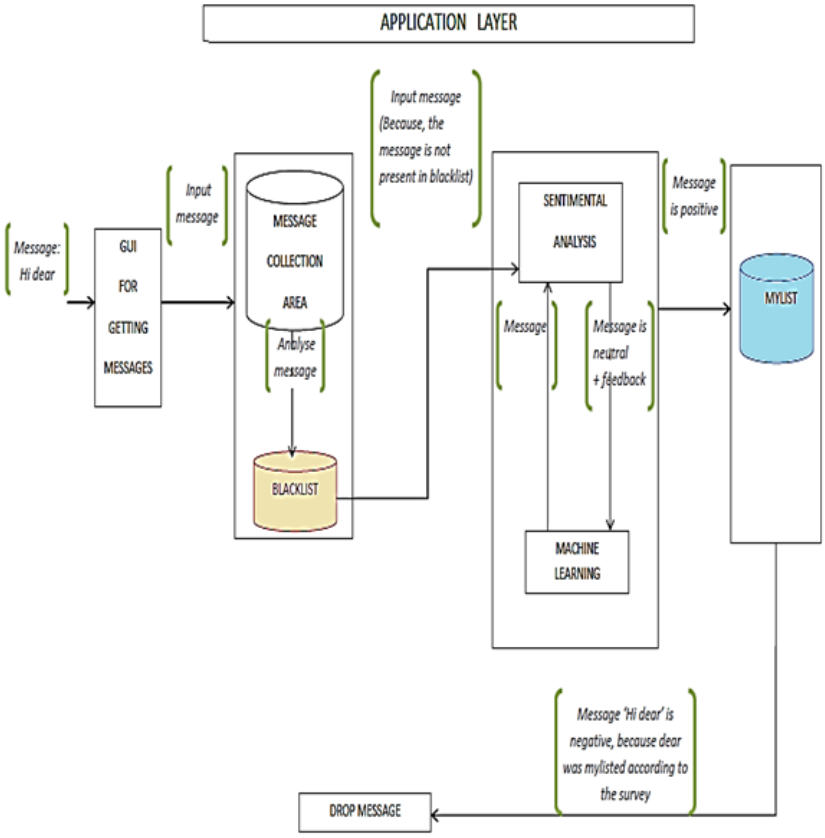

Figure 7- Working model of sample input which contains mylist words.

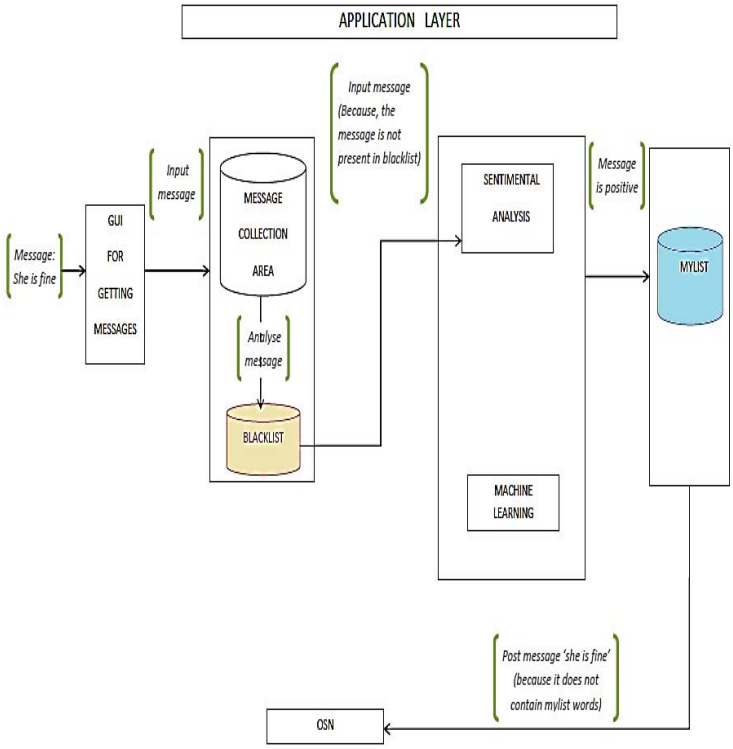

Figure 8 - Sample input which does not have mylist words.

First, the system requests the user to input the message that need to be posted. The received message is initially processed with the words in blacklist. If it contains any blacklisted words then the message will be dropped.Then the output is given for sentimental analysis using Stanford Core NLP for further analysis of tweets which classifies thesentence into positive, negative and neutral. The negative words are given to theblacklist and the messages are dropped. If it is neutral, the feedback is collected from the user about the tweet and it is again given back to NLP for analysis. Further analyzing the word with mylisted words to check whether it contains any customized bad words. If any, then the message is dropped.

3.2 FUNCTIONING OF SECURITY LAYER

The analyzed message post will be digitally signed using Blind signature schemes, which allows a person to get a message signed by an authorized party without revealing any information about the message to the authorized party. On approval of the authorized party the message is posted on the social networking site. Messages that cross the initial filters and not approved by the authorized party gets dropped at this layer.

3.2.1 BLIND SIGNATURE ALGORITHM

1. RSA blind signature scheme

TABLE2. TABLE OF DESCRIPTIONS

SYMBOLS	DESCRIPTION
m	Message
(n, e)	Public key of authority
d	Private key of authority
r	Random number
x	Product of message and blind factor
s’	Blinding signature

Algorithm 1.1 specifies the steps carried out in blind signature

ALGORITHM 1.1 BLIND SIGNATURE

INPUT – Message.
OUTPUT – Digitally signed signature.

STEP 1 – Start
STEP 2 – Obtain public key of authority (n, e).
STEP 3 – Choose a random number r, such that gcd (r, n) =1.
STEP 4 – Calculate x= (mr^e) mod n.
STEP 5 – Send the computed value of x to authority.
STEP 6 – Authority calculates and returns the signed values, s'= (x)^d mod n.
STEP 7 – Stop.

4. DISCUSSIONS

4.1 EXPERIMENTAL RESULTS OF APPLICATION LAYER

With the methodology explained in the previous section, a system has been implemented in java to test the capability of the system in securing message post. The observations made are discussed in this section:

1. An input message is given to a web application called ChatGaurd that facilitates safe texting. The messages are analyzed and filtered at four levels before they get posted in social media as shown in *figure 9*

Figure 9 Graphical user interface for analyzing tweets.

Level 1 – The message is analyzed to filter blacklist word. If it contains any offensive blacklist words then the message is automatically dropped else the message post is given to NLP for further analysis.
Level 2 - On passing the message to NLP tool, it takes the message as input and gives the corresponding sentiment such as positive, negative, neutral. The positive message post is given to next level for further analysis. The negative sentiment message post is automatically dropped. Whereas the neutral message post with the feedback is given back to NLP tool.

TABLE 3. LEVEL 2 RESULTS OF NLP TOOL

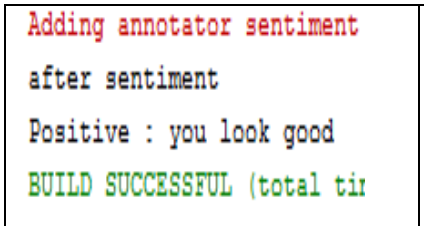

Figure 10 Example for positive sentiment message

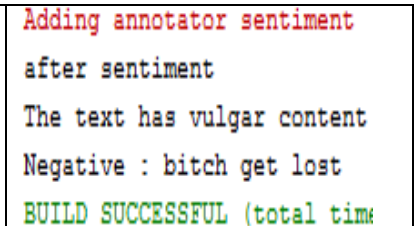

Figure 10.1 Example for negative sentiment message

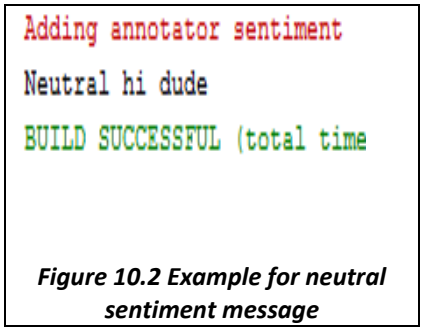

Figure 10.2 Example for neutral sentiment message

Level 3- The output of the NLP tool is further analyzedwith mylisted words to remove the customized bad words in the message. If the message contains any Mylisted words then the message is dropped else the message post is given to the security layer.

TABLE 4. LEVEL 3 RESULTS OF MYLIST

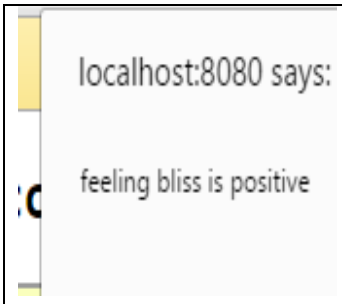

Figure 11 Example for positive message

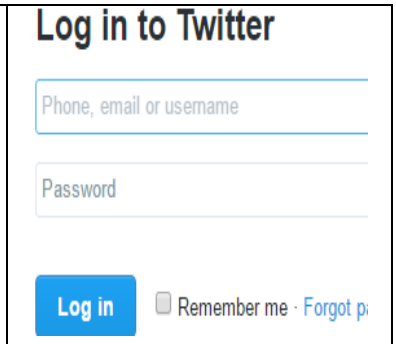

Figure 11.1 Redirecting to online social networking site if the message is positive

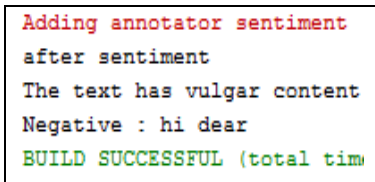

Figure 11.2 Example for negative message for filtering of Mylisted words(dear is Mylisted since it is considered as offensive by parents)

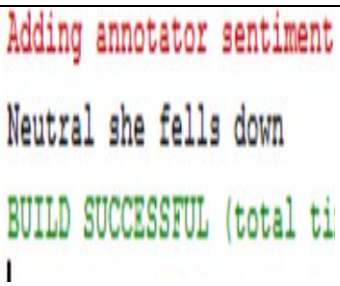

Figure 11.3 Example for neutral message

Level 4 – The output from application layer is given to security layer where the message is encrypted with blinding factor and the encrypted message is blindly signed by authorized party.

TABLE 5. LEVEL 4- SECURITY LAYER

The negative word present in the tweet is bitch THE TWEET IS go away you bitch The signature is from an authorized analyser.Th BUILD SUCCESSFUL (total time: 19 seconds)	THE TWEET IS i am happy The signature is not from an authorized analyser BUILD SUCCESSFUL (total time: 9 seconds)
Figure 12 Signature done and verified	***Figure 12.1 Signature done and not verified***

5. CONCLUSIONS

This paper provides an application for parents to have a control on the messages their wards post. Sentiment analysis of sentence of messages gives detailed results but developing a dictionary of offensive words makes it easy. The first step of the proposed scheme is to drop the unwanted words through blacklist mechanism. Next step is to classify the content and remove the customized words as per the wish of parents. This is followed by ensuring non repudiation without any compromise in privacy by the parent digitally signing the post using a blind signature scheme. On detailed analysis it was observed that the system has met all the objectives that are stated as contribution of the work under section 1.

6. REFERENCES

[1] Marco Vanetti, ElisabettaBinaghi, Elena Ferrari, Barbara Carminati, and Moreno Carullo. “A System to Filter Unwanted Messages from OSN User Walls”, IEEE Transaction on Knowledge and Data Engineering, Vol.25, No. 2, Febraury 2013.

[2] SaiSandeesh. R, Abishek. M, Harish. A. Raman, Karthik. V, Murali Krishna. E. “Analyzing the Twitter Feeds using Natural Language Processing and Machine Learning” , International Journal of Applied Engineering Research, Vol. 10, Issue 9, pp. 18911-18916,2015.

[3] U. Vanaja, M. Parthasaradhi. “An Efficient Rule Based System to Avoid Malicious Content from OSNs”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Special Issue 4, September 2014.

[4] V. Ezhilvani, K. Malathi, R. Neduchalian. “Rule Based Message Filtering and Blacklit Management for Online Social Network”. International Journal of Research in Engineering and Technology, Vol: 03 Issue: 06| Jun-2014.

[5] A. D. Swami, B. S. Khade. “A Text Based Filtering System for OSN User Walls”. International Journal of Advanced Research in Computer Science and Software Engineering, Vol 4, Issue 2, Febraury 2014.

[6] Amruta Kachole, S. D. Jondhale. “Unwanted Message Filtering System From OSNs User’s Wall using Customizable Filtering Rules and Black List Techniques”, (2014) International Journal of Emerging Technology and Advanced Engineering, 4(2), pp.752-755, ISSN (2250-2459).

[7] J. Anishya Rose, A. Pravin. “Machine Learning Text Categorization in OSN to Filter Unwanted Messages”. International Journal of Computer and Information Technologies, Vol. 5(1), 2014, 640-643.

[8] K.Babu, P.Charle. ”A System to Filter Unwanted Words Using Blacklists in Social Networks”, (2014) International Journal Of Computer Science and Information Technologies, 5(2), pp.1748-1753.

[9] Subhabrata Mukherjee, Pushpak Bhattacharyya. “Feature Specific Sentiment Analysis for Product Reviews”.

[10] N. J. Belkin and W. B Croft. “Information Filtering and Information Retrieval: Two Sides of the Same Coin?” Com. ACM, Vol. 35, no. 12, pp. 29-38, 1992.

[11] Saini Jacob Soman, Dr. S. Murugappan. “Detecting Malicious Tweets in Trending Topics using Clustering and Classification”, 2014 International Conference on Recent Trends in Information Technology.

[12] T. K. Das, D. P. Achariya, M. R. Patra. Opinion Mining about a Product by Analyzing Public Tweets in Twitter , International Conference on Computer Communication and Informatics (ICCCI), 2014 International Conference, pp.1-4, 3-5 Jan. 2014.

[13] Sunil Yadav, S Das, D Rudrapal. A System to Filter Unsolicited Texts from Social Learning Networks, Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference, pp.1-5, July. 2013.

[14] Bo Pang, Lillian Lee and ShivakumarVaithyanathan, “Thumbs up: Sentiment Classification using Machine Learning Techniques”, International Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002, pp. 79-86.

AUTHOR

SURYA PRABHA.K Student of Amrita School of Engineering, Ettimadai,Coimbatore, Department of Computer Science Engineering, (ksurya22m@gmail.com).

N HARINI Assistant professor at Amrita School of Engineering, Ettimadai, Coimbatore,Department of Computer Science Engineering,(n_harini@cb.amrita.edu).