ijecs
open access www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 10 October, 2014 Page No.8844-8849

Unwanted Message Filtering From OSN User Walls And Implementation Of Blacklist

Rakhi Bhardwaj[1], Vikram Kale[2], Prasad Morye[2], Manoj Dhaygude[2], Sagar Badhe[2]

[1] *Department of Computer Engineering, KJEI's Trinity College of Engineering & Research, Pune, India*

[2] *Department of Computer Engineering, KJEI's Trinity College of Engineering & Research, Pune, India*

ABSTRACT:

Today's Online Social Networks (OSNs) do not provide an ability to control the contents of the message posted on user's private wall. Any unwanted contents can be easily posted on the walls. Security in posting of unwanted message is an important issue in OSN. Up to now OSN have provided little control regarding who can post on user's private wall. Here we have proposed a system that will prevent unwanted messages from being posted on user's wall. To achieve this, the messages are scanned using text categorization techniques. We have also proposed a Blacklisting mechanism which will block such frequent message creators and prevent them from further posting any messages on the user's wall.

Keywords: Text filtering, text categorization, online social network, blacklist.

[1] Introduction

Online Social Networks (OSN) has provided vast number of people an easy communication medium. They have become today's one of the most popular interactive medium. OSN have helped people to keep in touch, share information about their daily activities, travels, photos and political upraising. Daily communication indicates the exchange of innumerous amount of data, text, images etc. The main part of OSN's is its public or private areas, called as walls. The contents displayed on these walls are generally visible to all the people in user's contact and sometimes also to the other people if the user allows it. OSN provides very less amount of security on the contents posted on these walls. Until today OSNs have provided a very little support to prevent unwanted messages on the user's wall. For example - Facebook, the leading social networking site, allows user to determine who can post on their walls (i.e. defined groups of users such as friends, friends of friends or anyone) but does not put any restriction over the contents to be posted. Messages are not scanned for vulgar, objectionable or political contents before being posted on user's wall. Therefore unwanted messages are easily posted on user's wall, no matter who posts it. Hence to overcome this problem, unwanted message filtering can be used. It will provide the OSN user walls with security against the posting of unwanted and undesired messages. It is one of the important requirements of an OSN service and is not provided so far.

The aim of the proposed system is to provide the user with a **Filtered Wall (FW)** mechanism. It is an automated system that will be able to filter out unwanted messages from social network's user wall. Also, the system will provide a **Blacklisting** mechanism. Blacklist will help user to block the people repeatedly posting unwanted content on former's wall. It can be achieved through text identification by scanning the message before posting it on a user's wall. Here first the text will be classified for unwanted contents using short text classifier technique, and then this text will be subdivided into different domains using text categorization. The next step is to determine the filtering rules and to perform the blacklisting management. Filtering is performed on basis of unwanted text or words in the message. Further the users who repeatedly try to post such unwanted contents will be blocked by our automated system and will be kept in blacklist for a certain period of time. The time period for which the user is being blocked or blacklisted depends on the contents of the messages posted by the user and the attempts made by him to post such messages. By using this technique OSN is provided with more security. This guarantees the prevention of circulation of undesired contents through online social networks.

[2] Related Work

Elena Ferrari, Elisabetta Binaghi, Marco Vanetti, Moreno Carullo and Barbara Carminati, [1] [2] have proposed a system that possess a mechanism to avoid unwanted messages from any user on an OSN wall of other users. This paper aims at providing OSN users an ability to secure their walls through filtering the unwanted contents being posted. This system will block the undesired messages sent by the user which is achieved by an automated system called Filtered wall (FW). Content based message filtering and short text classifier support this system. But the drawback of this system is that the user posting unwanted messages will not be blocked; only the message posted by the user will be blocked. To overcome this problem of the system, the term Blacklist will be implemented as future enhancement.

L. Roy and R.J. Mooney [3] have proposed a content-based book recommending system using information extraction and a machine-learning based algorithm for text categorization.

M. Demirbas, B. Sriram, D. Fuhry, E. Demir and H. Ferhatosmanoglu[4]. In this paper technique to classify messages on micro-blogging sites such as Twitter is explained. Messages on twitter are short and hence lack sufficient word occurrences. Therefore traditional classification methods such as "Bag-Of-Words" have limitations. Therefore, this paper proposes use of small set of domain-specific features extracted from the author's profile and text. This approach effectively classifies the text into sets of generic classes such as Private Messages, Opinions, Deals, Events and News.

[3] Existing System

Today's existing OSNs such as Facebook, MySpace, etc. provides a facility to user to allow the latter to choose a group of other users who can post the messages on latter's wall (i.e. defined groups of users such as friends, friends of friends or anyone). But, this provides little security to user's wall because the allowed people can post any kind of messages on the wall possessing unwanted contents like vulgar, objectionable or political words.

[3.1] Disadvantages of existing System

- The existing system does not scan the messages for unwanted contents before posting it on wall, no matter who posts it.
- It does not filter the messages possessing unwanted contents which the users don't want to be on their wall.
- It doesn't automatically block the people who keep posting unwanted messages on a user's private wall.

[4] Proposed System

In this paper we propose a technique known as filtered wall (FW), which is used for filtering unwanted messages. The Filtered Wall scans each message before being posted on wall. Filtering rules are used to determine which contents should be allowed on user's wall and which messages should be blocked. Further it will also provide a Blacklisting mechanism. Blacklist will be an automated mechanism which will block users posting undesired messages on the user walls. The prohibition can be approved for uncertain period of time.

The techniques used are:

- Filtered Wall
- Black Listing

i. Filtered Wall (FW): The first part of our system is the Filtered Wall (FW). In our proposed system each message being posted on OSN is scanned first by the automated Filtered Wall to detect if any unwanted contents are present. If any undesired contents are found the message is not posted on the wall with a warning to user. The user is well notified for the undesired contents in his/her message. The message is scanned using following techniques:

a. Short text classifiers: It is generally used when the amount of data to be classified is less. The main purpose of short text classifier is to identify the neutral words and categorize them from the non-neutral words. It should be done in step-by-step manner. In first step of classification the neutral and non-neutral data is separated. In the second step the classifier works on non-neutral data. For each of the non-neutral data it produces estimated appropriateness or grades. Such a list of grades is used by subsequent phase of filtering process.

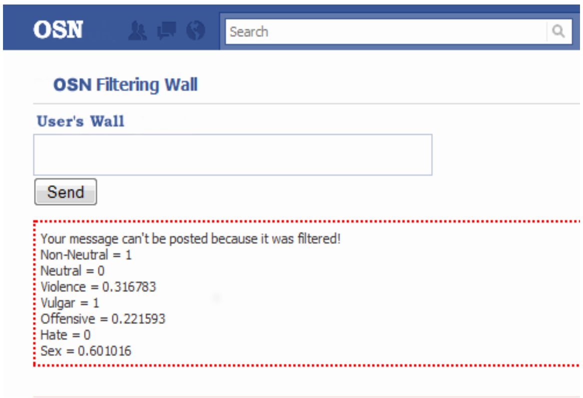
b. Text Categorization Techniques: The non-neutral words are further categorized into specific domains. The words in each domain are similar in nature. Each domain will have its own priority depending on the impact of the contained words. Some of the domains used are as follows:

- **Terrorism**: This domain will contain words related to terrorism like bomb, blast, terrorists, etc.

- **Criminal**: This domain will contain the words related to criminal intentions like murder, kill, attack, smuggle, rape, etc.
- **Vulgar**: This domain will contain words showing vulgarity like sex, fuck, etc.
- **Political**: This domain will contain words that may harm social stability which may result into communal riots, defamation, disputes among people.

These domains are previously created by the Network Admin. The Network Admin can edit these domains. In our proposed system we also allow the user to add a new domain and words to it if he/she doesn't want those words to be posted on their wall.

The expected result is as follows:



Your message can't be posted because it was filtered!
Non-Neutral = 1
Neutral = 0
Violence = 0.316783
Vulgar = 1
Offensive = 0.221593
Hate = 0
Sex = 0.601016

Figure: 1. Expected Result

ii. Blacklisting: This feature is an automated system which performs the function of blocking the user who repeatedly tries to post undesired messages on user walls. The user will be blocked depending on the impact of his messages and also the number of attempts the user makes to post such messages on the wall. The impact of these undesired messages is calculated considering the following factors:

- The total number of non-neutral words with respect to total words in message.
- The priority of the domains of non-neutral words.

The users will be blocked temporarily or permanently depending on the undesired contents of his/her message and the number of attempts the user makes to post such undesired messages on a user's wall. The Blacklist will be only accessible to the Network Admin who will have control over the list of blocked user.

[5] System Architecture

The above diagram represents the system architecture of our proposed system. Filtered wall provide the basic functionality of posting messages on OSN. It is generally supported by the filtering mechanism.

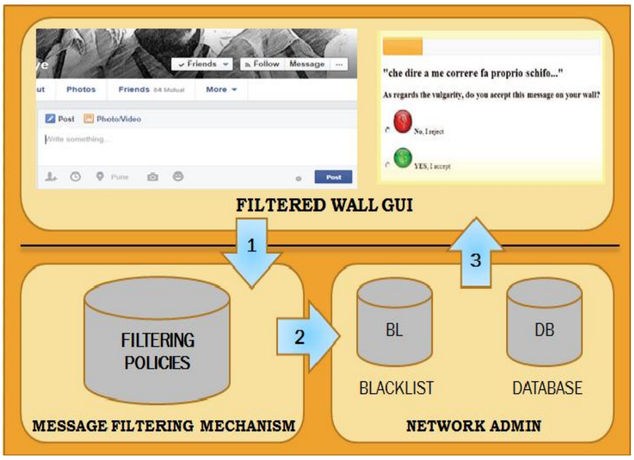


Figure: 2. System Architecture

The filtering mechanism provides the facility to filter each and every message posted on wall. After filtering the contents an acknowledgement is sent to network admin. The network admin is an automated system that performs the function of notifying the user for unwanted contents as well as performing the task of blacklisting. The contents of message are classified with the predefined set of unwanted words stored in data dictionary. If any unwanted contents are found then the user trying to post the message is notified. If the user repeatedly tries to post such unwanted messages on OSN wall then he/she is blacklisted from the OSN. The time period for which the user is been blocked is determined by an automated mechanism. The network admin maintains the list of blocked users separately.

[6] Conclusion

Inspecting the messages posted on OSN user walls is important issue in today's world. Our proposed system uses an automated mechanism to scan the messages before being posted on the user's wall and further filters those messages from OSN user walls which are unwanted and undesired. We have also proposed an automated Blacklisting mechanism which blocks the users who repeatedly try to post such undesired messages ignoring the given warnings. Hence, our proposed system provides more security to OSN user walls and therefore no objectionable or undesired contents can be circulated through our proposed mechanism for OSN user walls.

References

[1] Elena Ferrari, Elisabetta Binaghi, Marco Vanetti, Moreno Carullo and Barbara Carminati, "A System to Filter Unwanted Messages from OSN User Walls", IEEE Transaction, Vol. 25,No. 2,Feb 2013.

[2] Elena Ferrari, Elisabetta Binaghi, Marco Vanetti, Moreno Carullo and Barbara Carminati, "Content-based Filtering in On-line Social Networks", IEEE Transaction, Vol. 25,No. 2,Feb 2013.

[3] Raymond J. Mooney, Loriene Roy, "Content-Based Book Recommending Using Learning for Text Categorization".

[4] M. Demirbas, B. Sriram, D. Fuhry, E. Demir and H. Ferhatosmanoglu, "Short Text Classification in Twitter to Improve Information Filtering".

[5] Christian Bizer, Richard Cyganiak, "Quality-driven information filtering using the WIQA policy framework", Web Semantics: Science, Services and Agents on the World Wide Web 7 (2009) 1–10.

Authors:

i. Rakhi Bhardwaj:

She is an assistant professor at Computer Department of KJEI's Trinity College and Engineering and Research, Pune.

Address: KJEI's Trinity College of engineering and research, Pisoli, Pune, India. Pin Code-411048.

ii. Vikram Kale:

He is student of Computer Science at KJEI's Trinity College and Engineering and Research, Pune.

Address: KJEI's Trinity College of engineering and research, Pisoli, Pune, India. Pin Code-411048.

iii. Prasad Morye:

He is student of Computer Science at KJEI's Trinity College and Engineering and Research, Pune.

Address: KJEI's Trinity College of engineering and research, Pisoli, Pune, India. Pin Code-411048.

iv. Sagar Badhe:

He is student of Computer Science at KJEI's Trinity College and Engineering and Research, Pune.

Address: KJEI’s Trinity College of engineering and research, Pisoli, Pune, India. Pin Code-411048.

v. **Manoj Dhaygude:**

He is student of Computer Science at KJEI’s Trinity College and Engineering and Research, Pune.

Address: KJEI’s Trinity College of engineering and research, Pisoli, Pune, India. Pin Code-411048.