Filtering Mechanism for Unwanted Messages on OSN User Wall Using CBMF (Content Based Message Filtering)

Aishwarya Kashyup[1], Ravina Kothari[2], Himanshu Kothari[3], Geetika Narang[4]

Department of Computer Engineering,
Savitribai Phule Pune University,
Sinhgad Institute of Technology, Lonavala

***Abstract* - In this era of internet, the Online Social Networks (OSNs) are the platform to build social relations among people who share interests, activities, backgrounds or real life connections. OSNs have gained a ubiquitous status and this has led to security issues of posting unwanted messages on user wall. Therefore, in order to make the OSN user wall a secured wall, we are introducing a flexible- rule based system which provides users to control the messages that are posted on their walls and allows user to customise the filtering criteria to be applied on their walls. This system exploits machine learning based soft classifier for automatically labelling messages in support of content based filtering.**

Keywords- Online Social Networks, Flexible-rule based system, filtering criteria, Machine – learning based soft classifier, content based filtering.

I. INTRODUCTION

The Online Social networks have become a necessity in our daily life. The OSNs are a popular medium of interaction where large amounts of human information are disseminated. Several contents such as audio, video, texts, images etc are shared on a daily basis. The statistics of facebook says, an average user creates 90 pieces of contents each month, whereas more than 30 billion pieces of content are shared each month. This has led to the breach of privacy on OSN user wall and as a consequence the user might face the issue of unwanted messages on their walls.

Therefore in order to prevent the unwanted messages to be posted on OSN user wall information filtering is used. Thus, the proposed system uses a flexible rule based filtering that allows user to customise the filtering criteria that is to be applied to their walls.

This service is not only a matter of using web content mining techniques for different application, rather it requires to design additional classification strategies and also wall post comprises of short text and traditional methods have limitations. Here the short texts are separated and categorised based on its content using machine learning text categorization that automatically assigns short texts to a set of categories.

The motive of this work is to propose an automated system called Filtered Wall (FW). This system provides the user the privilege to control the messages that are posted on their walls.

We make use of neural learning as far as learning model is concerned. This is one of the best solutions in text classification. We base the overall short text classification strategy on RBFN for their proven capabilities in acting as soft classifier.

Our model implements a hierarchical two level classification strategy. In the first level the RBFN classifies the messages into neutral and non-neutral. In the second level, the non-neutral messages are classified producing gradual estimates of appropriateness to each category. After the classification is done, the filtering rules are applied.

Filtering rules give the result of ML categorisation process, which filter the user wall and relationship of user. In addition, the system provides the support for user defined black lists (BLs) that is lists of users that are temporarily prevented to post any kind of messages on OSN user wall.

The two fundamental concepts are Content Based Filtering and Short Text Classifier as explained below:

A. Content Based Filtering

An Information Filtering system is a system that eliminates the redundant or unwanted information from a stream of dynamically generated dispatched asynchronously by an information producer and present to the user those information that are likely to satisfy his/her requirements. It derives the information from the correlation between the information contents and the user preferences. The documents processed in content based filtering are somewhat similar to text classification. It exploits Machine Learning paradigm according to which a classifier is automatically induced by learning from a set of pre-classified examples.

B. Short Text Classifier

The short Text Classifier is a technique used for classifying short texts. It is based on neutral learning strategy to semantically categorise short texts. From ML point of view, this task requires a hierarchical two-level strategy. In the first level it classifies the sentences as

neutral and non-neutral. Then it classifies and eliminates the neutral sentences. In the second level classifier acts on non-neutral short texts and for each of them it simply produces estimated appropriateness or gradual membership for each of the conceived classes, without taking any hard decision on any of them. Such a list of grades is then used by the subsequent phases of the filtering process.

Let us take a look at the example as follows:

TABLE I

LEVEL OF CLASSIFICATION

I LEVEL CLASSIFIER	II LEVEL CLASSIFIER
Health	Health Poison
Boy	Bad Boy
Issue	Political issue

II. SYSTEM ARCHITECTURE

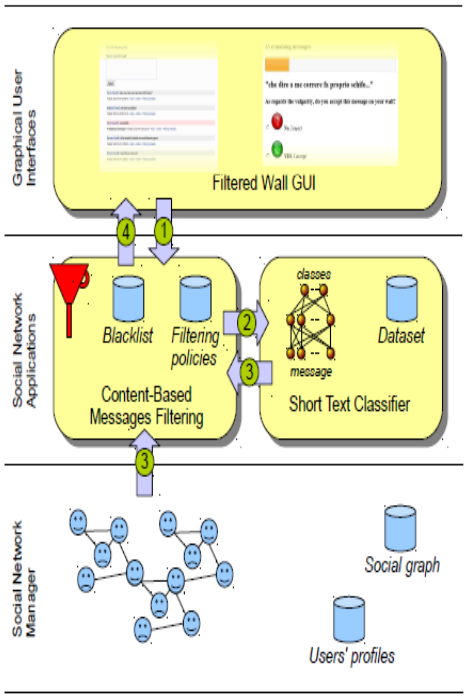

Fig.1 System Architecture

The osn architecture is three-tier architecture. The first layer of the architecture is Social Network Manager (SNM). The aim of SNM is to provide basic OSN functionalities (i.e. profile and relationship management). The second layer provides the support for external Social Network Applications (SNAs). SNAs may in turn require an additional layer for graphical user interface (GUIs). According to the architecture the proposed system is placed in the second and third layers. The users interact through GUIs, which provides users with FW, a wall where messages authorised by FRs/BLs are published. The fundamental components of the proposed system are Content-Based Message Filtering and Short Text Classifier (STC) modules. Then the latter classifies the messages into different categories. CBMF exploits the messages classified by STC in order to enforce FRs specified by the user. BLs can also be added to improve the filtering process. Given below is the summarised working of the given architecture:

1) After entering the private wall of one of his/her contacts, the user tries to post a message, which is intercepted by FW.
2) A ML-based text classifier extracts metadata from the content of the message.
3) FW uses metadata provided by the classifier, together with data extracted from the social graph and users' profiles, to enforce the filtering and BL rules.
4) Depending on the result of the previous step, the message will be published or filtered by FW.

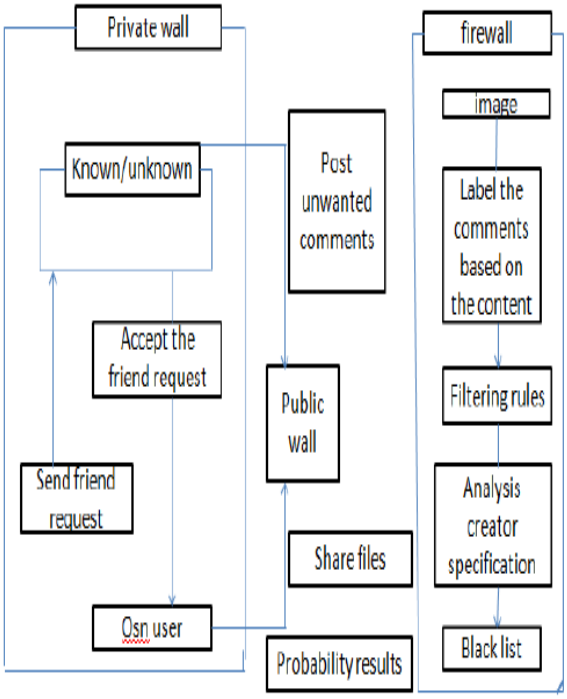

Fig 2. The flow of working of the Filtered wall

III. FILTERING RULES AND BLACKLISTS

Here we are going to describe and elaborate the filtering rules and then we illustrate the blacklists. We are going to map the social network to the directed graph in which each node corresponds to the user whereas the edges denote the relationship between the nodes. The edges represent the trust level between two users, that is, how much a given user considers the trustworthy with respect to that specific kind of relationship the user with whom he/she is establishing the relationship.
We assume that the trust levels are the rational numbers in the range [0, 1]. Let us say, there exists a direct relationship of the given type RT and trust value X between two users, if there is an edge connecting them having the labels RT and X. moreover, two users are in an

indirect relationship of a given type RT if there is a path of more than one edge connecting them, such that all the edges in the path have label RT. There are many algorithms to compute trust values. Such algorithms mainly differ on the criteria to select the paths on which trust computation should be based, when many paths of the same type exist between two users.

A. Filtering Rules

We consider three main different issues in defining the language for FRs specification that should affect the message filtering decision. When we talk about OSNs then it may happen that the same message may have different meaning and relevance based on who writes it. Therefore FRs must give the users the privilege to put constraints on message creators.

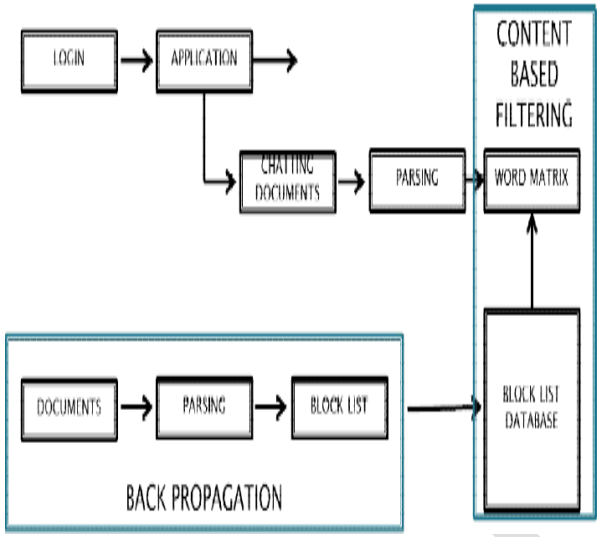

Fig 3. Architecture of the Filtered wall

There are various criteria through which message creators can be identified. There are the options formalised to specify the creators, defined as follows:

Definition 1. (Creator specification). A creator specification creator Spec implicitly denotes a set of OSN users. It can have one of the following forms, possibly combined:

1) a set of attribute constraints of the form an OP *av*, where *an* is a user profile attribute name, *av* and OP are, respectively, a profile attribute value and a comparison operator, compatible with *an*'s domain.

2) a set of relationship constraints of the form (*m*; *rt*;*minDepth*; *maxTrust*), denoting all the OSN users participating with user *m* in a relationship of type *rt*, having a depth greater than or equal to *minDepth*, and a trust value less than or equal to *maxTrust*.

Definition 2. (*Filtering rule*). A filtering rule FR is a tuple (*Author, creatorSpec, contentSpec, action*), where:

- author is the user who specifies the rule;
- *creatorSpec* is a creator specification, specified according to Definition 1;
- *contentSpec* is a Boolean expression defined on content constraints of the form (*C, ml*), where *C* is a class of the first or second level and *ml* is the minimum membership level threshold required for class *C* to make the constraint satisfied;
- *Action* € {*block; notify*} denotes the action to be performed by the system on the messages matching *contentSpec* and created by users identified by *creatorSpec.*

B. Blacklists

BLs is a mechanism to prevent messages from unwanted users. They are directly managed by the system through a set of rules called BL rules. BL rules allow the users to identify and block the unwanted users according to their profile as well as their relationships in the OSN. The BL rules gives the user the privilege to proscribe the user from their wall, they do not know directly or any user they know and have a bad opinion about them. The users are proscribed for a limited time span. The BL rules are formally defined as follows:

Definition 3. (BL rule). A BL rule is a tuple (*Author, creatorSpec, creatorBehavior, T*), where:

- Author is the OSN user who specifies the rule, i.e., the wall owner;
- *creatorSpec* is a creator specification, specified according to Definition 1;
- *creatorBehavior* consists of two components *RFBlocked* and *minBanned*. *RFBlocked* = (*RF, mode, window*) is defined such that:
  - RF = $\frac{\#bMessages}{\#tMessages}$, where *#tMessages* is the total number of messages that each OSN user Identified by *creatorSpec* has tried to publish in the author wall (*mode = myWall*) or in all the OSN walls (*mode = SN*); whereas *#bMessages* is the number of messages among those in *#tMessages* that have been blocked;
  - *window* is the time interval of creation of those messages that have to be considered for RF computation;

minBanned = (*min, mode, window*), where *min* is the minimum number of times in the time interval specified in window that OSN users identified by *creatorSpec* have to be inserted into the BL due to BL rules specified by author wall (mode = myWall) or all OSN users (mode = SN) in order to satisfy the constraint.

- T denotes the time period the users identified by *creatorSpec* and *creatorBehavior* have to be banned from author wall.

Example 3. The BL rule:
(*Alice; (Age < 16); (0:5; myWall; 1 week*); *3 days*)
inserts into the BL associated with *Alice's* wall those young users (*i.e., with age less than 16*) that in the last week have a relative frequency of blocked messages on *Alice's* wall greater than or equal to 0:5. Moreover, the rule specifies that these banned users have to stay in the BL for three days. If Alice adds the following component

(*3, SN, 1 week*) to the BL rule, she enlarges the set of banned users by inserting also the users that in the last week have been inserted at least three times into any OSN BL.

IV. EVALUATION

A system will automatically filter unwanted messages from OSN user walls on the basis of both message content and the message creator relationships and characteristics. The paper substantially extends for what concerns both the rule layer and the classification module. Major differences include, a different semantics for filtering rules to better fit the considered domain, an online setup assistant (OSA) to help users in FR specification, the extension of the set of features considered in the classification process, a more deep performance evaluation study and an update of the prototype implementation to reflect the changes made to the classification techniques. In web mining the most general sense it can contribute to the increase of profits, be it by actually selling more products or services, or by minimizing the costs. In order to do this, marketing intelligence is required. This intelligence can focus on marketing strategies and competitive analyses or on the relationship with the customers. The different kinds of web data that are somehow related to customers will then be categorized and clustered to build detailed customer profiles. This not only helps companies to retain current customers by being able to provide more personalized services, but it also contributes in the search for potential customers.

V. CONCLUSION

In this paper, we propose a system to filter unwanted messages from OSN walls. On-line Social Networks (OSNs) is to give users the ability to control the messages posted on their own private space to avoid any unwanted content to be displayed on the user wall. Up to now OSNs provide little support to this requirement. To fill the gap, we enhance the system by creating an instance randomly notifying a message system that should instead be blocked, or detecting modifications to profile attributes that have been made for the only purpose of defeating the filtering system. Additionally, we plan to investigate the use of online learning paradigms able to include label feedbacks from users. Moreover, the flexibility of the system in terms of filtering options is enhanced through the management of BLs. In particular, future plans contemplate deeper investigation on two interdependent tasks. The first task concerns the extraction and/or selection of contextual features that have been shown to have high discriminative power. The second task involves the learning phase. Since the underlying domain is dynamically changing, the collection of pre-classified data may not be representative in the longer term. The present batch learning strategy, based on the preliminary collection of the entire set of labelled data from experts, allowed an accurate experimental evaluation but needs to be evolved to include new operational requirements. In future work, we plan to address this problem by investigating the use of on-line learning paradigms able to include label feedbacks from users. Additionally, we plan to enhance our system with a more sophisticated approach to decide when a user should be inserted into a BL. The development of a GUI and a set of related tools to make easier BL and FR specification is also a direction we plan to investigate, since usability is a key requirement for such kind of applications. We would like to remark that the system proposed in this paper represents just the core set of functionalities needed to provide a sophisticated tool for OSN message filtering. Moreover, we are aware that a usable GUI could not be enough, representing only the first step. Indeed, the proposed system may suffer of problems similar to those encountered in the specification of OSN privacy settings. As future work, we intend to exploit similar techniques to infer BL rules and FRs. Additionally, we plan to study strategies and techniques limiting the inferences that a user can do on the enforced filtering rules with the aim of bypassing the filtering system, such as for instance randomly notifying a message that should instead be blocked, or detecting modifications to profile attributes that have been made for the only purpose of defeating the filtering system.

REFERENCES

[1] Ms. Shruti C. Belsare, Prof. R.R. Keole, "OSN user filtered walls for unwanted messages using content mining", IJCSMC, vol.3, issue 3, march-2014, pg 97-103.

[2] Dipali D. Vidhate, Ajay P. Thakare,"to avoid unwanted messages from osn user wall: content based filtering approach", IJCSMC, VOL.3, Issue.4, april 2014,pg 688-692.

[3] R. J. Mooney and L. Roy, "Content-based book recommending using learning for text categorization," in Proceedings of the Fifth ACM Conference on Digital Libraries New York: ACM Press, 2000, pp. 195–204.

[4] M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari, "Content-based filtering in on-line social networks," in Proceedings of ECML/PKDD Workshop on Privacy and Security issues in Data Mining and Machine Learning (2010)

[5] F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1–47, 2002.

[6] H. Sch¨utze, D. A. Hull, and J. O. Pedersen, "A comparison of classifiers and document representations for the routing problem," in Proceedings of the 18th Annual ACM/SIGIR Conference on Resea. Springer Verlag, 1995, pp. 229–237.

[7] R. E. Schapire and Y. Singer, "Boostexter: a boosting-based system for text categorization," Machine Learning, vol. 39, no. 2/3, pp. 135– 168, 2000.

[8] A. Adomavicius, G.and Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE Transaction on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734–749, 2005.

[9] 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, 2010, pp. 841–842.

[10] 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, 2008, pp. 1444–1445.

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

[12] M. Carullo, E. Binaghi, and I. Gallo, "An online document clustering technique for short web contents," Pattern Recognition Letters, vol. 30, pp. 870–876, July 2009.