Beyond Classical Filters: A System Where Blacklists are Obtained Using Filtering Rules

Ms.Suvarna Bahir[1], Mr.Syed A.H[2]
[1, 2] Aditya Engineering College,Beed.

Abstract- *It is quite common nowadays that people can use a social networking site to post any kind of message on a particular user's profile. These messages can be both vulgar and good messages and are read by everyone connected in the network. This paper proposed a system which filters unwanted messages and implements a blacklist using a flexible rule based system that allows a user to customize filtering criteria to be applied to their walls. The user can be temporarily blocked if his/her degree of vulgarity is high then he/she can directly add into the blacklist.*

Keywords- Information Filtering, Content Based Filtering, Short Text Classifier, Online Setup Assistant, Blacklist.

I. INTRODUCTION

Online Social Networks (OSNs) are the most preferred nowadays by each and every generation. It is the upcoming trend in people's life. The rapid growth of OSNs has significantly influenced human lifestyles, especially in the patterns of communication and cooperation. These OSNs provide a proper and effective way of communication means through the use of contents as such, texting, sending various images and videos or audio among the users. Since OSNs can provide a huge pool of manpower and support quick diffusion of information, they can be a basis for immediate and effective cooperation among people. These OSNs allow a large number of users who are globally dispersed to connect to each other, thus, providing a valuable opportunity to enhance the level of social cooperation towards achieving some common goals.

There are several kinds of messages been send over the network from place to place, so it becomes necessary to develop a classification mechanism, so that the data which is found useless by the users eliminate it. Therefore, the system will provide the users the automatic way of controlling the messages before being posted on their walls. Till date, OSNs actually do not provide much support to avoid the unwanted messages on the walls i.e. for friends, friends of friends, or a well defined group of people. Thus, it is not possible to avoid these undesired messages regarding who posts it to you. To detect such posts there is a great need of classification techniques, as there is traditional use of short text by users, which do not provide sufficient words that are understandable by the system.

Daily and continuous communications imply the exchange of several types of information, including free text, image, audio, and video data. It is quite common nowadays that people can use a social networking site to post any kind of message on a particular users profile i.e. their wall. These messages can be both vulgar and good messages. These messages are read by everyone connected in the network as the messages are posted on the walls of the user. For example, good messages can be, "Happy Birthday." and vulgar messages can be like, "Oh..!! sh*t." and so on. Therefore, the system designed thus emphasizes on these messages and filters them according to the Content-Based Message Filtering (CBMF)[5]technique which is based on neutral and non-neutral classification of messages. Also filtering rules (FRs) are setup using Online Setup Assistant (OSA), which helps in short text classification techniques [6].

The purpose of this paper is to build a system which works in an automated form without disturbing the user walls. Therefore, there is use of Filtered Wall (FW), which does half of the system work i.e. filtering of the offensive messages. The Machine Learning (ML)[1] strategies are also studied, that will assign a set category to each and every short text which can be better understood by the system. There is a major role played by the Short Text Classifier (STC) helping in re-building of short text that cannot be considered as good messages by the system. In this the words are firstly extracted so that classification can be done on them. For STC we require the exogenous knowledge related to the context posted that will enlarge the short text which are endogenous before been enlarged. Such short text will then go through the filtering rules, so that they do not come under the category of unwanted messages. The overall short text classification techniques are based upon the In particular, we base the overall short text classification Radial Basis Function Networks (RBFN) for enacting as soft classifiers for balancing the rogue material send. There are two levels in the classification strategy. The work of the RBFN is to classify the short messages as Neutral and Non-neutral messages in the second stage. These neutral messages are nothing but the nice or readable messages but the non-neutral messages are those which come under the

vulgar or the bad messages. Further to implement this RBFN there is a need to specify the Filtering Rules (FRs), that states which kind of text be or cannot be displayed on their walls. These FRs provide several type of filtering criteria that can be used by the users to benefit and reach their needs. Moreover, the FRs present the relationship between the users, their walls and also their profiles. Depending on the context the FRs are properly implemented accordingly and enforced on the message to be displayed on the wall. All those above functionalities are firstly implemented on the Facebook as one of the Online Social Networking site, but it can be also applied on other OSNs as well.

The system to be designed consists of complete working of Blacklists (BLs) which shows the users being blocked and setting up user-defined Filter Rules using OSA tools. It also assigns the trust values to each user independent of any kind of relationships with each other and assigns it according to their behavior in the OSNs. The user can be temporarily blocked if his/her behavior is not proper in terms of degree of vulgarity i.e. if the degree is high then the users are directly added into the blacklist by giving him/her a prior notice about the immature behavior he/she shows. The consideration firstly is given to the extraction and/or selection of contextual features which have been portrayed having high discriminative power. Then the concerned part in the next stage involves the learning phase. There is a need of development of a GUI (Graphical User Interface). And also a set of related tools to make easier BL and FR specification, since usability is an important requirement for such kind of applications. In particular, the system aims at investigating a tool able to automatically recommend trust values for those contacts user does not personally know. Therefore the proposed system consists of such a tool that should suggest trust value based on users actions, his/her reputation in the social network, which might implement enhancing OSNs with audit mechanisms and such a service is not yet been provided so far in the OSNs.

The remainder of this paper is organized as follows: Section 2 surveys related work, whereas Section 3 introduces the conceptual architecture of the proposed system. Section 4 describes the ML-based text classification method used to categorize text contents, whereas Section 5 illustrates FRs and BLs. Section 6 illustrates the performance evaluation of the proposed system, whereas the prototype application is described in Section 7. Finally, Section 8 concludes the paper.

II. RELATED WORK

The main contribution of this paper is implementation of Blacklist with customizable content-based message filtering [5] for OSNs, based on ML technique[2].To the best of our knowledge, this is the first proposal a system to automatically filter unwanted messages from OSN user walls on the basis of both message content and the message creator relationships and characteristics with blacklist. The current paper substantially extends [4] 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.

1. Content Based Message Filtering

In this topic, we introduce you to the principles we have adopted for filtering the unwanted messages and texts. The definition of the language for filtering rules specified states, three main aspect that, from our perspective, should affect the filtering decision. The first issue relates to the fact that, in Online Social Networks like in routine life, the message that appear to be same must have variant meanings and connection based on its written by whom. Consequently, users should be allowed by the filtering rules to state restrictions on message creators. Thus, several different criteria must be set on the basis of which the creators must be selected on which a filtering rule applies, one of the most important is to impose conditions on user profile's characteristics. In this way, for example, it is possible to state rules that apply only to young/teenage creators, to them with a given religious/political background, or to those creators that we think are not expertise within a specified given field. In the given OSNs, another way by which the creators can be recognized, is with the help of their social graph on which information can be exploited. This suggests to define conditions on the 3 aspects depth, type and trust values of the relationship[6]. Also the creators should be participated so that the specified rules are applied on them. The next relevant aspect to be taken into consideration in describing the language for precise filtering rules is the emphasis on and support for content-based principles. This states that the filtering rules recognize messages on the basis of restrictions on its contents. So as to state and emphasize these constraints, the 2-level text classification that is introduced later is used. To be more precise, the plan is to develop class of the first and second level and also their membership level so that the users become capable of defining content-based constraint. For instance, the messages with high probability would be recognized, as to neutral or non-neutral, (i.e., the Neutral/Non-Neutral first level class messages related with membership

level are higher than the stated threshold); also, in the same manner, with a specified second level class.

Other aspect to be believed, is deserved to be taken into account is dealing with the problems and obstacles that on an average OSN user face in stating the right threshold for membership level. To the user's comfort level in specifying the membership level threshold, our opinion is that it would be helpful to allow the specification of a bearing value that is related with every base constraint, defines how much the membership level can be made lower than the membership threshold in that constraint. Also introduction to the tolerance value would be useful for the system to manage, the messages in proximity to the rule and so they are worth a special way of treatment. Particularly, these messages are the messages that has membership level threshold greater than the membership level intended in the filtering rule but equal or greater to the defined tolerance value[5].

For instance, we might have a principle that will need obstruction in or to block messages those are with violence class that has membership level which is greater than 0.8. So in this case messages that has violence class with a membership level of 0.79 are displayed, because they are not filtered as stated in the rule. But if we introduce a tolerance value of about 0.05 in the previous content-based constraint then it permits the system or it will automatically handle and manage such kind of messages[1]. The complexity arrives when the system has to react with messages those found only for the tolerance value and this is a complicated issue to be dealing with many variant ideas. Because of its complex nature and, more priorily, there is a requirement of an thorough experimental assessment, in this paper we have adopted a new resolution according to which a notification will be sent to the user by the system related to the messages to ask his/her opinion. So we push ahead the investigation of these ideas to be done in future. The last and final aspect of the filtering rule is the act that is to be performed by the system on the messages that fulfill the needs of the rule. The “block”, “publish” and “notify” are the possible actions that we have considered, with relating to the meaning of blocking/displaying the text, or send notification to the user about the message so to wait for his/her opinion.

2. Policy-Based Personalization of OSN Contents

Recently, proposals exploiting classification techniques have been found for personal access in OS networks. For instance, a classification method is been developed to classify short text messages so that we can prevent excited users that use microblogging services with the help of data that is raw. The system that is been described in the emphasize on Twitter2 and relates a set of types with each tweet defining its matter. The user can then see just certain category of tweets that are based on her/his interest. To show contrast, Kuter and Golbeck developed an application, that is called "FilmTrust", that explains provenance information and OSN trust relationships to make a personal access to website. However, filtering policy layer are not provided by such systems because of which the result of the classification process can be exploited by the user, so that the user can decide how or to which extent we can filter out unwanted data and messages. In contrast, our filtering policy language also permits the setting of filtering rules based on a variety of criteria, that will take into account not only the results of the classification method but also wall owner relationships with another OSN users also information written on the user profile[1]. In addition, our system is blessed with a flexible feature mechanism that is used for BL management process that will integrate a further chance of customization to the filtering way of processing. The OSN service that we are offering that is providing filtering capabilities to its users is the "MyWOT", http://www.mywot.com is a social networking service which provides its subscribers the two main facilities as given:

1) Resources that rate the system based on respect to four criteria those are: vendor reliability, trustworthiness, child safety, and privacy;
2) Preferences that are defined that determine if the browser should or not obstruct access to given resource, or that it should simply warn by sending a message which is based on the specific type of rating.
3) Besides the existence of some same things, the way adopted by the MyWOT is different from that of ours. Particularly, filtering criteria is supported which is far less flexible than those ones of that of the Wall that is filtered as they are only related on the four criteria above mentioned.

In addition, the end user is not provided with the automatic classification mechanism. Many access control models are inspiring our work and enforcement mechanisms related policy languages that have been developed so far for OSNs, since several similarities are shared by filtering with the help of access control. In the environments of OSNs, access control models that are developed so far force access control that is topology-based, based on which the access control needs are explained in terms of relationships that resource owner with the requester should have. We use a similar strategy to recognize the users to whom a FR can be applied. Dealing to filter unwanted matter and not with access control, the availability of a described message contents one of the key ingredients of our system is that is to be explained by

the filtering mechanism process. In oppose, previously cited no one of the access control models exploited the content contained in the resources so as to force access control. More detailed, the above-mentioned access control models do not consider the notion of BLs and their management. Lastly, the policy language is also related with the policy frameworks that have been developed so far to support enforcement of policies and the specification that are expressed in terms of restrictions on the resource that is understandable by the machine descriptions of them that are provided by Semantic web languages. It gives the ability to use filtering policy so that we can denote the ”quality” needs to the end users that are to be satisfied by the web resources in order to be displayed to the users. However, even if such frameworks are very strong and general enough to be customized or/and are extended for variant application scenes they are not been particularly viewed to report the filtering of information in OSNs and. And hence, we select to state our abstract and more compressed policy languages, instead of extending one from those mentioned above.

III. FILTERED WALL ARCHITECTURE

The architecture of filtered wall consists of main three layers. The very first layer of the architecture is Social network manager (SNM) which commonly provides the basic OSN functionalities. The next second layer is social network application (SNAs) which provides support for external social network applications.3rd layer is Graphical user interfaces (GUIs)which provide user with filtered wall i.e. wall where only authorized messages are published according to blacklist and FRs. Our proposed system is placed in 2nd and 3rd layers. Core component of this proposed system are CBMF and STC.

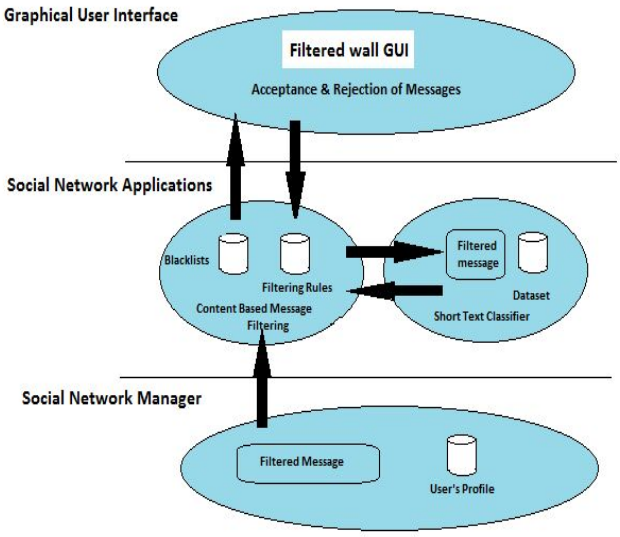

Figure 1. Filtered wall conceptual architecture and the flow messages follow from writing to the publication.

This is Filtered wall architecture which tells us how the message will flow and how that message will be published on the wall. For successful filtering of unwanted messages from the OS/n user walls certain following steps are to be achieved. In 1st step when one user who is a direct or an indirect user of another OSN user enters on his/her private wall, and tries to post a message is very oftenly always intercepted by FW. Filtered wall filters the message according to its content i.e. whether the text to be posted on the desired wall is good or bad verbally. Now in CBMF, blacklist contains the list of blocked users and filtering policies allows the setting of FRs according to variety of criteria, which consider classification of messages and user profile and their relationship. Short text classifier divide messages into number of classes such as neutral, non-neutral, vulgar, violence etc. Database contains metadata, text classifier extract metadata from the content of messages. In 3rd step 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. Finally in the last step analysis of the result is done and decide whether message will be published on the wall or not, which is done by the Filtered Wall architecture process and further be discarded automatically. Following these rules it is decide whether a message is posted on the wall or filtered by the filtering rules.

IV. DICOM FW

A new prototype DicomFW is been introduced, it is Facebook application whose work is to imitate a user's personal wall, so that the user can apply a specific set of combined proposed FRs. Throughout the overall development of the DicomFW prototype our emphasis is to give attention just on the FRs, leaving the Black List (BL) mechanism implementation for future work. As the implemented functionality it permits the STC and CBMF components to interact with each other it is a critical functionality. The contextual information (information from which CF can be extracted) relevant to the name of the group are not accessible directly because the application is considered as wall and not a group.

Figure 2. DicomFW: A message filtered by the wall's owner FRs (messages in the screenshot have been translated to make them understandable).

Prototype uses contextual information currently is associated to the name of the group where the user is most active by whom the message is written. So that the prototype is extended in the future, we want to add all contextual information which is related to all the group name's in which the user is involved, weighted appropriately by the level of participation. It is very important to note that this type of contextual information is associated to the environment to which the user prefers to post the message, thus the experience of trying using DicomFW is constant with what is defined and evaluated in the Section of Experimental Evaluation. So we can summarize, our application allow to:

1) view the specific list of users' FWs;
2) view messages and post a new one from them on a user FW;
3) state FRs using the OSA (Online Setup Assistant) tool.

By this when a user will try to post some message on a wall, then he/she will be sent a notification telling if they are blocked by the FW or whether the message will be actually be displayed on the user's wall. .

V. SHORT TEXT CLASSIFIER

The techniques those are established used for classifying text are likely to work well on datasets which contains large documents like as newswires corpora, but they will suffer if the documents in the corpus are likely to be short. The defined set of discriminant and characterizing features are the critical components in this context and they allow the representation of concepts underlying and also the representation of collection of a consistent and complete set of supervised instances. A multi-class soft classification method is the work of semantically classifying short texts and that includes two important phases: to represent text and classification based on ML. The study is aimed at evaluating and designing several representation techniques in a combined approach with the neural learning strategy to meaningfully categorize short texts[3]. Through the Machine Learning point of view, we take in the task by defining a hierarchical two-level strategy assuming that it is better to eliminate after identification "neutral" sentences, then classify "non-neutral" sentences by the class of interest instead of performing everything in a single step. The first-level task is taken into account as a hard classification in which short texts are labeled with static Neutral and Non-neutral labels. The second-level soft classifier acts on these crisp set of non-neutral short texts and, for each one of them, it "simply" evaluates estimated appropriateness or "gradual membership" for each of the conceived classes, without performing any "hard" decision on any one of them. This list is then used by the consequent stages of the filtering mechanism[1].

A greedy strategy approach is used to select the feature set, it follows the definitions of classes. It extracts 8 features (8F) , consisting of one nominal (author) and seven binary features (presence of shortening of words and slangs, time-event phrases, opinioned words, emphasis on words, currency and percentage signs). Categorization of messages into the selected classes requires the knowledge of the source of information. Hence, the authorship information selection is the primary feature. Suppose that we define an event as "something happens at a given place and a particular time", the presence of participant, place, and time information shows the existence of an event in the text[6]. Hence, the date/time information and time-event phrases which are extracted and collected from a set of messages based on general observation of users and set it as a feature. The user information is also taken into consideration and stored with the presence of the '@' character which is further followed by a username within the messages in twitter. The emphasis on the words which are based on the use of uppercase are also captured. One more way is repetitive use of characters in a word (e.g., gooood). Twitter allows its users to send private messages by making the use of character '@', which is followed by username at the beginning of the message. Thus, the system captures the private messages by "@username", at the beginning of posts.

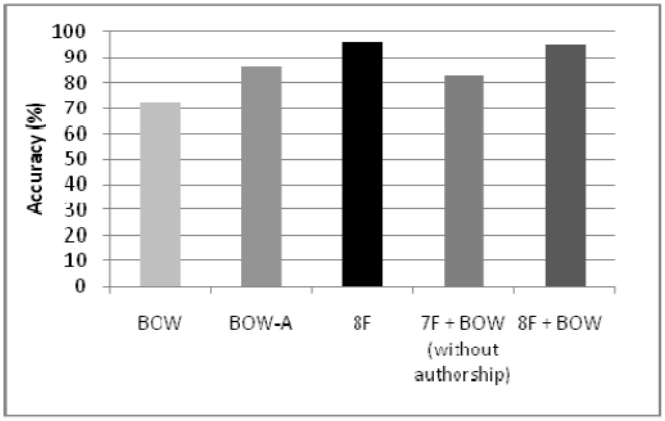

Figure 3. Overall accuracies.

In Figures, BoW refers to Bag of Words, BoW-A is BoW with the author feature,, and 8F represents our approach. As shown in Fig, on the overall accuracy 8F achieves 32.1% improvement over BoW. It is found that the author feature is differential in the provided dataset. BoW-A achieves 18.3% improvement over BoW, and even 3.7% over 7F+BoW (without authorship) on the overall accuracy. As shown in Fig 2, 8F performs consistently better for all classes, which can be used with BoW to have a better accuracy with an additional time cost of initial training. 8F achieves 35.2%, 103.4%, 12.2%, 9.9%, and 87.0% improvements over BoW for N, O, D, E, and PM, respectively[6]. In BoW, misclassified tweets are mainly between N and PM (383), N and O (407), whereas in 8F, they are mainly between N and O (104). We attribute this to the fact that tweets in N may also be opinionated. We believe that multi-label classification would resolve this issue to a certain extent. The times taken to build the training models are 37.2 and 0.8 sec for BoW and 8F, respectively. The ratio between these timings will be larger with larger collections as the number of words (features in BoW) will increase, while the feature count in 8F stays fixed.

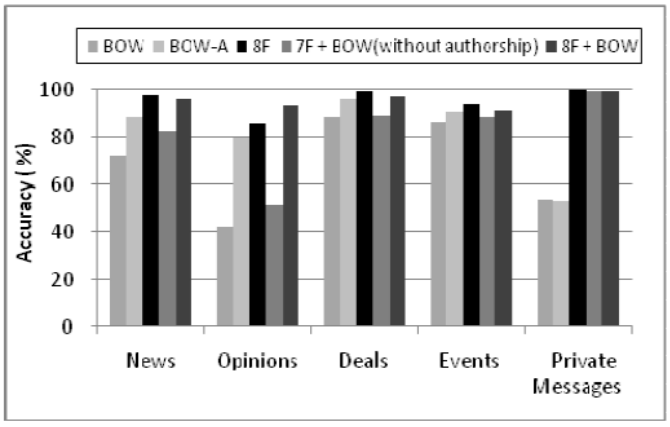

Figure 4. Accuracies for individual classes.

1. Text Representation

This extraction is a crucial and important task intensively affecting the performance of the overall classification strategy.

Text categorization and retrieval tasks are often based on a good representation of textual data. Different sets of features for text categorization have been proposed in the literature [4].On the basis of our experience we consider three contextual features, BoW, Document properties (Dp) and contextual features (CF). The first two features i.e. BoW & Dp are endogenous implies that they are entirely derived from the data contained within the text of the message. Such kind of text representation has a good general applicability, though in their operational settings, it is compulsory to even use exogenous knowledge implying any source of information other than the text in the message body but somehow linked to the message itself. Thus our system uses CF modeling information which categorizes the posting environment.

These endogenous and exogenous features play a key role in understanding the message semantics[4].All these features are evaluated and analyzed in the subsequent phases in order to determine the combination best suited for short text classification.

VI. ONLINE SETUP ASSISTANT

Online Setup Assistant (OSA) is an assistant tool for providing the threshold values to all the messages tried to be posted on the Filtered Wall. Online Setup Assistant is a set of procedures in which a well defined group of series of certain type of questions are asked to the user which is selected from the dataset. In this OSA procedure the user tells the system about his/her final decision whether to accept the message or reject the message. The type of messages that are accepted or rejected are studied by the system and the person who does not send a message within those criteria set by the user are asked to be inserted into the blacklists automatically without any interventions. According to the collection of the messages that are accepted or rejected are properly maintained and analysed by the system computers, which shows the users attitude towards what kind of messages or posts the user wishes to see on the wall. But for this entire selection and rejection process certain procedure has to be followed[1].

For example: Hypothecate that Bob is an OSN user for any OSN and is not interested to watch any kind of posts that has a high degree of vulgarity and nevertheless, also wants to block the another user for the same. Therefore, by following the OSA procedure for the above, the Vulgar class according to the user's need is set to 0.8. Also further the user wants to filter such degree of offensive messages only from

the indirect people going through his profile, while for the individuals with whom the user is quite familiar he wishes to block only those messages whose trust value is below 0.5. This process can be successfully achieved through the following presented FRs:

- ((Bob; friendOf; 2; 1), (V ulgar; 0:80), block)
- ((Bob; friendOf; 1; 0:5), (V ulgar; 0:80), block)

Consider any one friend of Bob who has a trust value of about 0.6 and he tries to post a message on the Bob's wall which has a certain degree of vulgarity. The message suppose is, "G*d d*mn f**k*ng s*n of a b*tch!" When the message is being posted, the message will be graded with a membership value of 0.85for the Vulgar class i.e. the degree of vulgarity of the message is shown here with a specified value with the message. Therefore, any kind of message having exceeded a certain limit of vulgarity set value will be successfully filtered from the system and as a result will not appear on the users wall as well.

VII. BLACKLISTS

An extended approach of our system is BL mechanism to avoid messages from ascertained creators, independent of their contents. These Blacklists are directly managed by the system, which should be able to predict the users to be inserted into the BL and decide when they should be removed from the BL i.e their retention in the BL is done. This is done using BL rules, to enhance flexibility. These rules are not made by the SNMP, rather they are specified by the users themselves that who all has to be banned from posting on the owners wall and for how long. However, if a user is banned from posting on one persons wall, he might be at the same time allowed to post on other persons wall. Similar to the filtering rules the BL rules make the wall owners to identify which users to be blocked based on their relationships in the OSNs and their profiles. So, with the advent of BL rules, wall owners are, for example, able to ban users they do not know personally from their walls (i.e. with them they have only indirect relationships). To make out the persons behavior in the OSN, we have taken into consideration two criteria. The first being that if particular user has been inserted into the BL for Several amount of time in a given span, say greater than the given limit or threshold, he/she might deserve to stay in the BL for some more time as their actions have not improved. Secondly, in contrast to catch the new intolerable behaviours, we use the Relative Frequency (RF) which allows the system to be able detect the users whose messages are in a constant habit to fail the FRs. These can be considered either locally or globally[1].

A BL rule is statically defined as: A BL rule is a tuple ðauthor, creatorSpec, creatorBehavior, TÞ, 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 ¼ (RF, mode, window) is defined such that

1. RF ¼ #bMessage ,#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 ¼ myWall) or in all the OSN walls (mode ¼ SN); whereas #bMessages is the number of messages among those in #tMessages that have been blocked.
2. 2)Window is the time interval of creation of those messages that have to be considered for RF computation; minBanned ¼ (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 ¼ myWall) or all OSN users (mode ¼ 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:

ðAlice; ðAge < 16Þ; ð0:5;myWall; 1 weekÞ; 3 daysÞ

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.

VIII. EXPERIMENTAL EVALUTION

To provide an overall estimation of how strongly the system applies a FR, we can again view at Table 1[1]. This table will permit us to assess the Recall and Precision of the FRs, since the values that are reported in Table 2[1] are been computed for FRs with the content specification component fixed to be (C,0.5). Let us assume that the given rule is applied

by the system on a specific message. So as such that, exact values reported in Table 2 is the probability that the opinion taken on the message that is taken into account (i.e, may be blocking it or not) is the correct one actually. In oppose to it, Recall has to be explained as the probability that, a given rule must be applied over a message, that really has to be enforced. We can now have a look and discuss about the results those are presented in Table 2, which reports Recall and Precision values. The Recall and the Precision value calculated for FRs are represented in the second column of Table 2 with (Neutral, 0.5) constraints. With the contrast, the fifth column represents the Precision and the Recall value calculated for FRs with (Vulgar, 0.5) content constraint. Results obtained by renowned information filtering methods are not so good as compared to those obtained by the content-based specification aspect, which are on the first level classification, which are good reasonably and enough aligned results achieved for the content-based specification aspect for the second level classification are less intelligent than those achieved for the first level classification,

Table 1. Results for the Two Stages of the Proposed Hierarchical Classifier

Text Representation		First Level Classification		Second Level Classification		
Features	BoW TW	OA	K	P	R	F_1
Dp	-	69.9%	21.6%	37%	29%	33%
BoW	binary	72.9%	28.8%	69%	36%	48%
BoW	tf-idf	73.8%	30.0%	75%	38%	50%
BoW+Dp	binary	73.8%	30.0%	73%	38%	50%
BoW+Dp	tf-idf	75.7%	35.0%	74%	37%	49%
BoW+CF	binary	78.7%	46.5%	74%	58%	65%
BoW+CF	tf-idf	79.4%	46.4%	71%	54%	61%
BoW+CF+Dp	binary	79.1%	48.3%	74%	57%	64%
BoW+CF+Dp	tf-idf	80.0%	48.1%	76%	59%	66%

Table 2. Results of the Proposed Model in Term of Precision (P),Recall (R), and F-Measure ðF1Þ Values for Each Class

	First level		Second Level				
Metric	Neutral	Non-Neutral	Violence	Vulgar	Offensive	Hate	Sex
P	81%	77%	82%	62%	82%	65%	88%
R	93%	50%	46%	49%	67%	39%	91%
F_1	87%	61%	59%	55%	74%	49%	89%

Table 3. Agreement between Five Experts on Message Neutrality

	Classification		Neutral			Non-Neutral		
Expert	OA	K	P	R	F_1	P	R	F_1
Expert 1	93%	84%	97%	93%	95%	97%	93%	95%
Expert 2	92%	80%	91%	98%	94%	95%	78%	85%
Expert 3	95%	90%	99%	94%	97%	88%	99%	93%
Expert 4	90%	76%	89%	98%	93%	94%	73%	82%
Expert 5	94%	84%	94%	97%	95%	93%	85%	89%

But it is necessary for us to explain this in view of the intrinsic problems in giving it to a messages a semantically most particular type. However, as we have analysed the features written in Table I which shows that improvement in the ability of the classifier so that it can properly differentiate between non nuetral classes is due to the introduction of contextual information (CF). Because of this all policies are made more reliable thoe exploiting non-neutral classes, which forms majority in todays real-world scenes.estimation of the k value can be done with the training sets of various fractions from TABLE 4[1].

Table 4. Agreement between Five Experts on Non-neutral Classes Identification

	Violence			Vulgar			Offensive			Hate			Sexual		
Expert	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Expert 1	89%	99%	94%	89%	97%	93%	80%	90%	85%	78%	98%	87%	82%	98%	89%
Expert 2	77%	83%	80%	92%	67%	78%	71%	60%	65%	71%	69%	70%	85%	67%	75%
Expert 3	81%	84%	83%	76%	96%	85%	67%	79%	72%	53%	89%	66%	84%	76%	80%
Expert 4	96%	41%	58%	92%	78%	84%	70%	60%	65%	79%	42%	54%	97%	64%	77%
Expert 5	84%	90%	87%	92%	77%	84%	77%	73%	75%	78%	84%	81%	85%	77%	82%

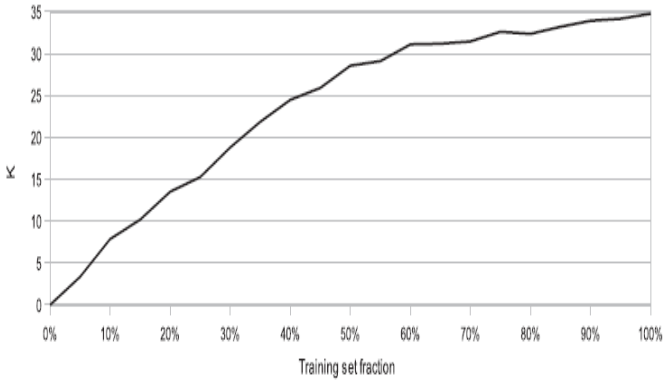

Figure 3. K value obtained training the model with different fractions of the original training set.

IX. CONCLUSION

A system to filter unwanted and unnecessary messages from OSN user walls has been discussed in this paper. The system makes the proper and effective use of FRs, ML mechanism, CBMF, STC, Policy Based Personalisation rules and implementing Black Lists as well. The implementation of BLs have made the system to be more efficient. The BLs are used to block the persons who oftenly show their undesired behaviours in terms of posting the messages to others. This therefore, increases the quality of the system and benefiting the OSN users from having a protective user wall.

Further, there is also much more enhancement in the system that decides itself when a user must be inserted into the BL, by giving a prior notification to the user who is misbehaving with some other user's wall, that is based on his/her actions in the OSN. In the first agenda, the wall posts are accordingly classified as per the type of posted message i.e., whether they are offensive or not. In the second agenda, it will classify those and discard the vulgar posts and only include those ones which are not offensive in nature. For classification of the various send texts there is a proper database of the offensive and defensive words, quotes, punctuations. This classification process is carried on the basis of CBMF technique. There is need of a developed GUI which will help one make proper use of the BL and the FRs, in order to have a user friendly system in accordance with the user. Therefore, a tool that automatically assigns the trust values to the user that are personally known or not known is given depending on their performance, actions, reputation and behaviour in the Online Social Networking site at present. For assigning this trust values to different user a audit mechanism is required to be setup to examine and supervise everyone present in the network. Inspite of this fact, the design of these audit-based tools is complicated by several issues, like the implications an audit system might have on users privacy and/or the limitations on what it is possible to audit in current OSNs. Even if we have complemented our system with an online assistant to set FR thresholds, the development of a complete system easily usable by average OSN users is a vast topic which was out of the scope in the earlier paper, which is not the case with the present research. As such, the developed Facebook application is to be meant as a proof-of-concepts of the system core functionalities, rather than a fully developed system.

However, it is found even today that the average OSN user finds it difficult to follow the general privacy settings and apply on their own account. Therefore, in our system, we intend to exploit these techniques to surmise BL rules and FRs. Additionally, we have studied techniques which limit the inferences that a user is able to perform on the enforced filtering rules, by successfully undergoing through a thorough filtering rule mechanism, where the notifying message is given to the misbehaving user, instead of being blocked. But later on after his/her several attempts to post these offensive messages he/she is blocked for set number of days. Repeating the same behavior increases their time span of remaining in the Black list. Moreover, our system can be easily implemented in any OSN and prove to be user friendly in case of an average OSN user. Though we have tried to put in various advanced functionalities alongwith the core ones, our system may suffer with some problems which are common in the privacy settings of any OSN. This is in scope to be diminished in our future work resulting in very advanced and best suited system for the users to keep their walls and profiles secure and protected from the ascertained users and the users with low trust values provided by the OSN on the basis of their behavior in it.

REFERENCES

[1] Marco Vanetti, Elisabetta Binaghi, Elena Ferrari, Barbara Carminati, and Moreno Carullo, " A System to Filter Unwanted Messages from OSN User Walls", IEEE Trans. on Knowledge and Data Eng.,, vol. 25, no. 2, Feb 2013.

[2] Zhi Xu and Sencun Zhu," Filtering Offensive Language in Online Communities using Grammatical Relations.", Department of Computer Science and Engineering. The Pennsylvania State University, University Park, PA 16802, szhu@cse.psu.edu

[3] 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.

[4] Gergely K´otyuk and Levente Butty´an, "A Machine Learning Based Approach for Predicting Undisclosed Attributes in Social Networks.". Laboratory of Cryptography and Systems Security (CrySyS) Budapest University of Technology and Economics.

[5] M. Vanetti, E. Binaghi, B. Carminati, M. Carullo and E. Ferrari, "Content-based Filtering in On-line Social Networks", Department of Computer Science and Communication University of Insubria 21100 Varese, Italy.

[6] Bharath Sriram, David Fuhry, Engin Demir, Hakan Ferhatosmanoglu Murat Demirbas," Short Text

Classification in Twitter to Improve Information Filtering.", Computer Science and Engineering Department, Ohio State University, Columbus, OH 43210, USA and Computer Science and Engineering Department, University at Buffalo, SUNY, NY 14260, USA.