Filtering Unwanted Messages from OSN User Wall Using Rule Based Techniques

[1]Mr. Akshay Bagal, [2] Prof. Shriniwas Gadage,

[1]Master of Engineering , [2] Assistant Professor
[1]Computer Engineering ,
[1] G.H.R.C.E.M, Wagholi, Pune,India

***Abstract* - The Online Social Network Provides a Platform to build Social networks and relationships among the people who wants to share information, news, interests and day to-day activities through the popular medium to communicate with each other. On OSN anyone can posts messages on their walls, where sometime peoples posts undesired or unwanted messages on users wall, so to prevent these unwanted messages. The system work is developed to prevent unwanted messages from osn walls to control the offensive, vulgar messages to be not posted on wall. The short text classification and text representation techniques are used to classify the contents of wall posts to categorize the messages. Using filtering rules on categorized messages, and different filtering criteria by defining categories of messages to be filtered and prevents the user from posting on wall. On the basis of these methodologies system develops a secure architecture to prevent the unwanted messages from OSN user wall and more flexible customizable system for efficient filtering techniques.**

***Index Terms* - Information Filtering (IF), Short text Classification, Filtering Rules (FRs), Content based filtering, offensive languages.**

I. INTRODUCTION

The rapid growth of social media to the social networking sites provides people to communicate with their friends, family members to share their views and lot of time they are haunting on social sites. As on these growth basis people cannot imagine their life without internet to the social sites. to express their emotions, news and activities of their private life on social sites. So according to Facebook survey billions of people use the social sites to share their activities in texts, images, videos where number of peoples share to thousands of activities per day. According to this the privacy to the social users is to be concern to provide more security to peoples to maintain their relations and personal profiles to be secured and clean. The users nowadays facing the problems related to the offensive languages like vulgar, bad, sexual and hate types of languages to posted on users wall which is unwanted and offensive to the users using social sites. The posts like offensive and political comments which is posted on users walls creates misunderstanding between public and to the private life of users so the users wants to be preventing from posts that are posted on walls.

For Filtering unwanted messages contains offensive languages that should be classified using text classification and extraction to the texts that occur in the posts. Short text classification and categorization is used to classify the content to find out the languages of text. Content based filtering provides the major part to filter the messages on the basis of content to which after filtering rules can be assigned to prevent or filter unwanted messages from user walls.

II. RELATED WORK

Marco Vanetti [1] firstly proposed the system to filter unwanted messages from OSN user wall on the basis of message content and the relationship of particular users within social network. And presents the idea of filtering rules and classification techniques to be applied on the messages posted on walls which prevents the unwanted messages. Introducing the short text classification works on the data sets which having large documents to be classified and evaluates different representation techniques. M.Chau [2] To find relevant data are very complicated on web content so it presents approach to web page filtering on the basis of content and structure analysis. Web pages consist of content based and link based feature in proposed system. Neural network approach is used for system to avoid useless data to be filtered and relevant data is discovered from large web content. R.J.Mooney [12] presents recommendation system which has methods like collaborative and content based recommendation which provides the find relation among the people. Collaborative filtering method which is the system that chooses items based on the correlation between people with similar preferences, but content based recommending system that develop information extraction and machine learning algorithm for text classifications. In this paper [7] Performance of classification includes different semantics for filtering rules is considered to be done. And provides the system to take decision about the messages which has to block or prevented from wall using content filtering, which depends on preclassified data set.

This paper presents [4] approach is ML text categorization technique, which automatically assigns each short text message from a set of categories based on its content provides soft classification techniques and performance of classification of messages to the neutral and non-neutral message categorization. A.Adomavicius [3] recommender system0s use three approach content-based recommendation, collaborative and hybrid recommendation. By Using this approaches we can enlarge recommendation system using contextual features of texts present in the large web contents. The policy personalization online social contents B.Sriram [5] in online services like twitter, categorizes the specific interest of users to the specific feature where classification is applied on the small set of categories consists of domain specific features where tweets can define its content and it helps to understand the interest of users. V.Bobicev [6] presents the consistent precision of partial matching of short text contents of specific terminology mention in the classification of texts. This model helps to provide specific terminology to fin the partial matching of texts classifications. J. Golbeck[8] proposes trust relationship like film trust application which exploits the relationship to the social network where user can trust on the ratings and reviews of film on the basis of criteria trustworthiness, privacy, vendor reliability and safety to the users privacy concerns. This way we provide the privacy for the users on the social network to prevent the private data of users. So proposed approach effectively categorize texts of a predefined set of classes which consists of opinions of users, news and activities held by the user on the social networks. The survey helps to understanding the concepts and models to be implemented in the development of the system to develop new approaches for flexible use of the system.

III. Problem Definition

In this section, we formally extracting and categorizing the messages which are posted on wall. The number of messages to be categorized where the filtering rules are applied on the categorized messages (i.e hate, offensive, violence, sexual, vulgar). Different weights are assigned on different messages according to category wise to classify that the messages is unwanted and cannot be posted on wall. If the unwanted messages are detected we can also block the user to post the messages on wall.

Rule based technique is used to find out unwanted messages and to prevent to be posted on wall.

IV. Proposed System Approach

Proposed system supports the user to give ability to control messages posted on the users wall, so first user login to the social sites then user posts the messages or text on friends wall. After posting text on friends wall system applies short text classifier to classify the text and extracts the texts from messages that posted on wall. Filtering rules are applied on the extracted texts from messages to check whether the messages is unwanted or not and particular message found to be unwanted then it is discarded and notification is send to the user who wants to post the message. The filtering criteria to recognize the content is unwanted or it is good to be decided by applying different rules. The categorization of messages are done on the category wise classification like vulgar, sexual, violence, offensive, hate where all available words are categorize on these categories. To find unwanted messages the words used in that messages are checked by classification and filtering rules whether the messages are found undesired then the weightage to that offensive word or any category that word belongs calculation is done according to it.

The unwanted words are determined from messages and category is assigned to unwanted words and weights are assigned to the words to find out that the messages posted on wall is objectionable or not on the basis of calculation of weights assign to words and to the category which it belongs if calculation is above threshold value then it is directly discarded. And if the value obtained from calculation of occurred words weights to the category is below threshold value then it is posted on users wall. In between the messages is found to be offensive or whatever category notification is given to the friend who wants to post the message.

Further the unwanted messages list are maintain to show the results recognized unwanted posts and black list is also created to list the different users who are blocked by the admin on the basis of the vulgar languages used in the messages.

Fig 1. Shows the way system implements the filtered wall to discard unwanted messages from users wall on the basis of content based message filtering and rule based approach.

It checks whether the user is already blacklisted or not if match is found directly notification is sent to the user and he cannot post the messages on wall. If user is not blacklisted then Filtering rules and classification techniques [4] are applied on the messages posted on the wall. If the content of message is relevant i.e.good then it is posted on wall and if message contains the irrelevant content i.e. Unwanted messages (vulgar, political,sexual, offensive etc.) then notification is sent to the user and messages is not posted on wall. In system the filtering criteria of messages is to be done on text categorization and extraction features with the help of Machine learning soft classification techniques [2], [4]. This way the system works and provides the privacy for the user which is achieved by implementing automated filtered wall [1]. Blacklist management helps to make decision on basis of users recent activities which maintain the list of blocked users and notifying by E-mail personally those who insist to post unwanted messages.

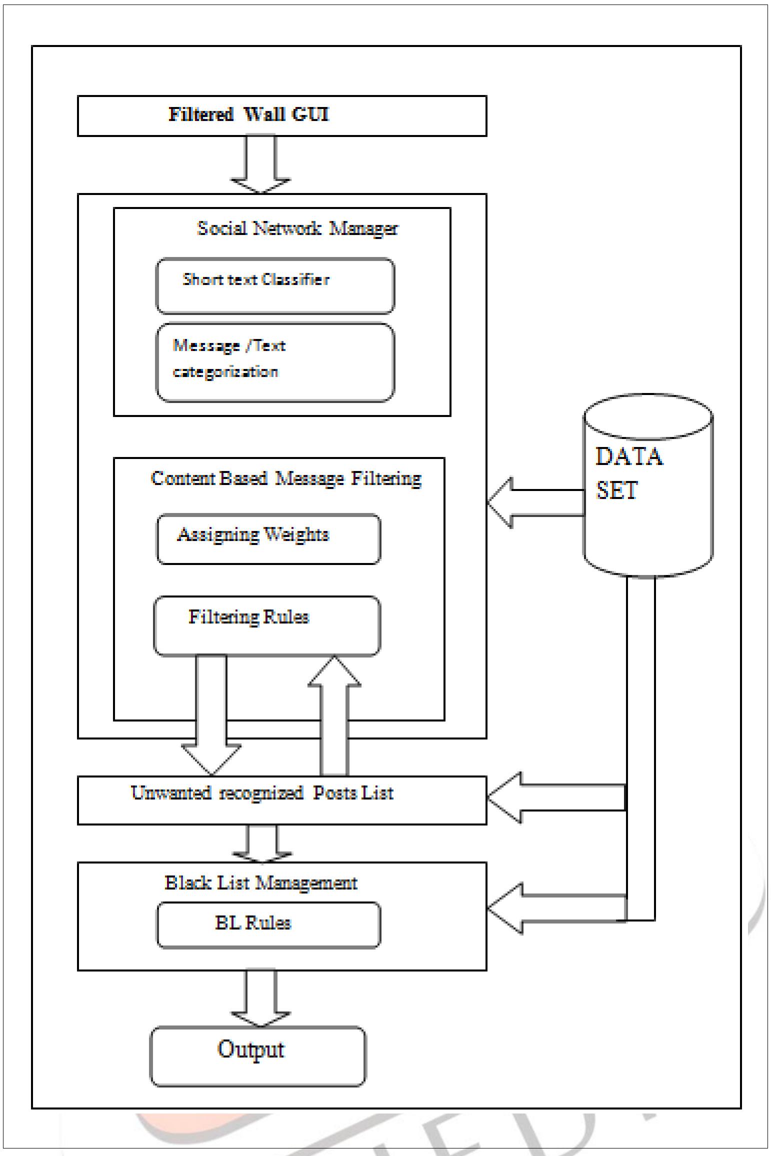


Figure 1: Architere of Proposed System

V. Methodology

Afterthetextedithasbeencompleted,thepaperisreadyforthetemplate.DuplicatethetemplatefilebyusingtheSaveAscommand,anduset henamingconventionprescribedbyyourconferenceforthenameofyourpaper.Inthisnewlycreatedfile,highlightallofthecontentsandimport yourpreparedtextfile.Youarenowreadytostyleyourpaper;usethescrolldownwindowontheleftoftheMSWordFormattingtoolbar.

Short Text Classification

Short text classifier characterizes classification on the small data sets and short texts. Our goal is designing and representing various discriminant features with a neural learning strategy which categorize short texts. A hierarchical two level classification strategy [13] is introduced for better to identify and eliminate "neutral" sentences and classifies "non-neutral" sentences by the class of particular interest, where the first level short texts are labeled as neutral or non-neutral and further second level non-neutral texts are classified.

Text Representation

Representation of text is an important task where the performance affecting the classification strategy is measured. the survey suggest three types of features considerations are[7], [9] Bag of words (BoW), Document Properties (Dp) and Contextual Features (CF) are used for text representations. First two types of features are entirely derived from the information contained within the text of message whereas contextual features are exogenous. Representing text using endogenous. Terms are identified with words in Bag of Words representation. It is also important to use Feature which is extracted from outside the message content but related to message itself. A contextual interests featured in that characterizes the environment where the user is posting According to Vector space Model is the model of text representation by which a text document is represented [11] as a vector of binary or real weights. These three features are experimentally evaluated for short text classification for their appropriateness.

Filtering Rules

Filtering rules are introduced to filter unwanted messages where users can state the rules for contents not to be displayed or blocked. Filtering rules are specified on the relationship of users and profiles. Filtering decisions are taken on these factors.

The rules are applied on the messages that found to be unwanted messages is categorize into five categories are Hate, Offensive, Violence, Sexual, Vulgar.

Rules:

Defines the occurrence of unwanted words are recognized with their distributed category. After recognizing words having their weights to the category count is taken if the particular category word count is above threshold value then it is discarded.

Likewise the rules are assign category wise, Priority basis the rules are applied on the messages.

Vulgar and Sexual Category- In vulgar category the rule is assigned that if the post contains vulgar word then it is totally discarded because its vulgar language contains high impact to the users profile.

Violence, offensive and hate- this three category of words contain less objectionable words that are abused on social network so the rule is applied on this categories is the weights assign to words to the category is calculated and defining certain threshold we can make decision whether to post the message or discard.

VI. EXPERIMENTAL RESULTS

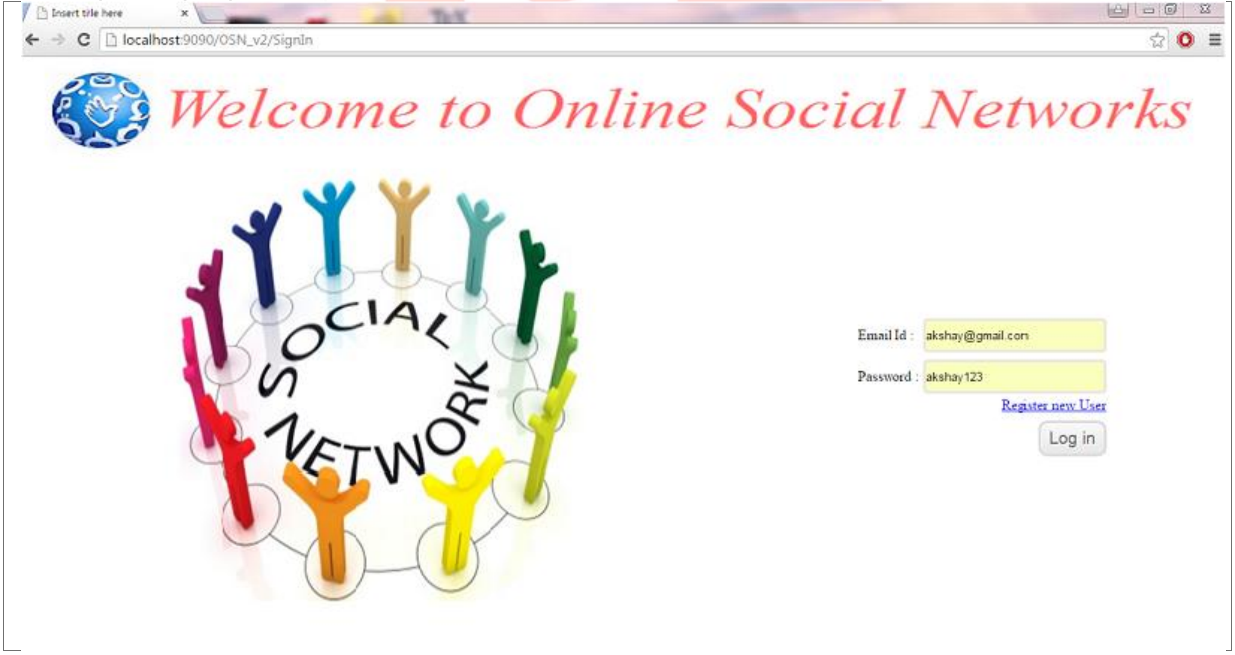
Input:

User posts the messages is the input data. This consists set of messages from different user to their friend walls posts. The messages are categorized in five categories (hate, offensive, violence, sexual, vulgar).

Outcomes:

Ouput generated b showing recognized unwanted posts to which it is prevented to be posting on wall. The difference between objectionable content to the non objectionable content is recognized and the results are shown on the list of unwanted posts which contains the users email id, unwanted content and the weights calculation on the occurrence of offensive words collection is shown to which the count is measured on the recognized words and category is assigned to each and every word which are detected to be in posts to find out that the post is unwanted. This is the way to which the unwanted content is recognized with the classification to the particular categorized within the category wise relational mapping between the occurred relevant to the ir-relevant data.

Following are the snapshots of a particular step wise execution where user interface defines how to interact system. The filtered unwanted words are shown in fig5 where we can see the results of a system. Finally the graph shows the system unwanted posts are filtered according to weightage.



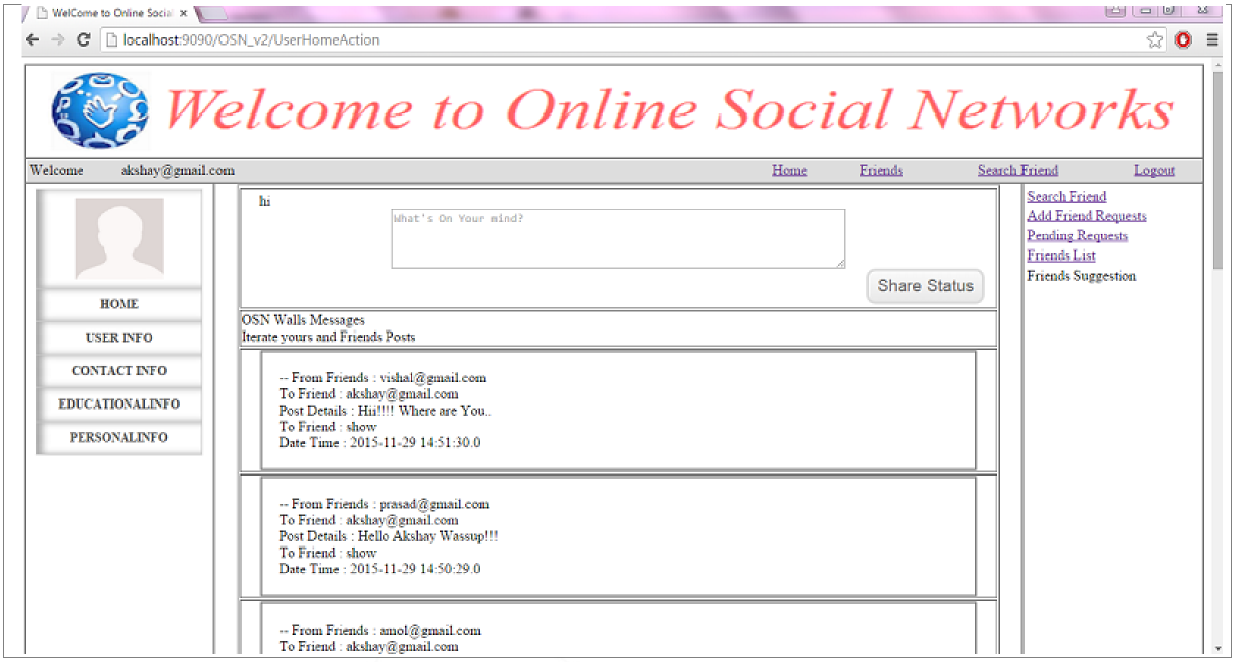
localhost:9090/OSN_v2/SignIn

Welcome to Online Social Networks

Email Id : akshay@gmail.com

Password : akshay123

Figure 2: Main Page of Login



Welcome akshay@gmail.com

hi

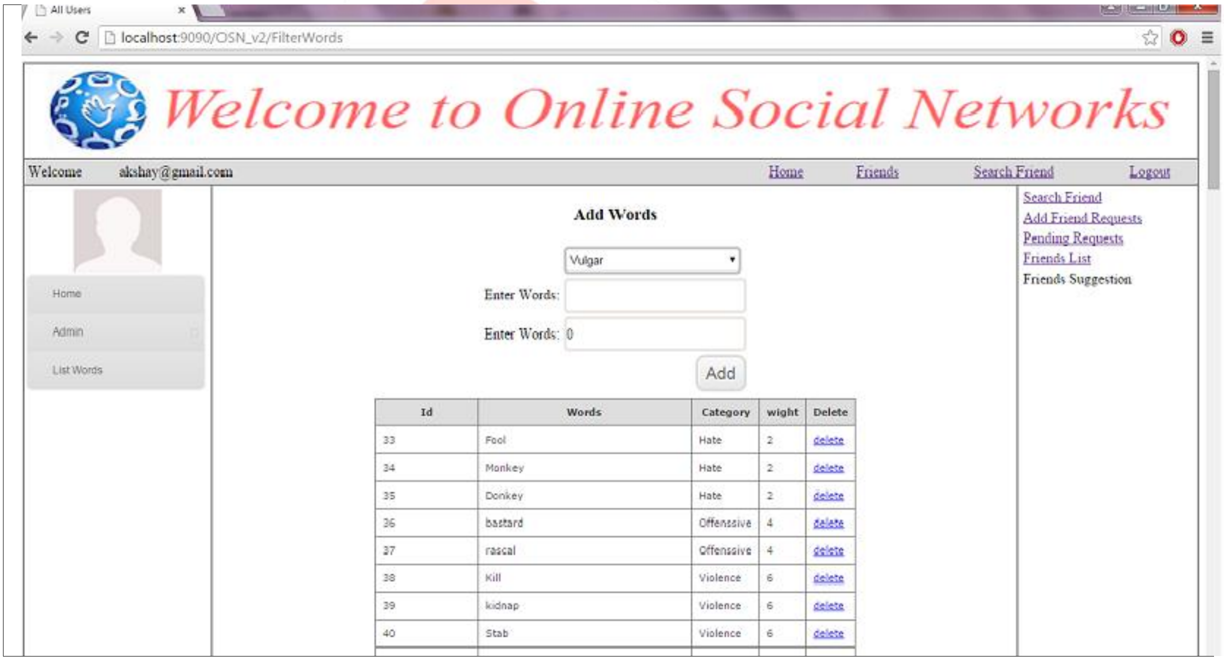
OSN Walls Messages
Iterate yours and Friends Posts

-- From Friends : vishal@gmail.com
To Friend : akshay@gmail.com
Post Details : Hii!!!! Where are You..
To Friend : show
Date Time : 2015-11-29 14:51:30.0

-- From Friends : prasad@gmail.com
To Friend : akshay@gmail.com
Post Details : Hello Akshay Wassup!!!
To Friend : show
Date Time : 2015-11-29 14:50:29.0

-- From Friends : amol@gmail.com
To Friend : akshay@gmail.com

Figure 3: User's Wall



Welcome akshay@gmail.com

Add Words

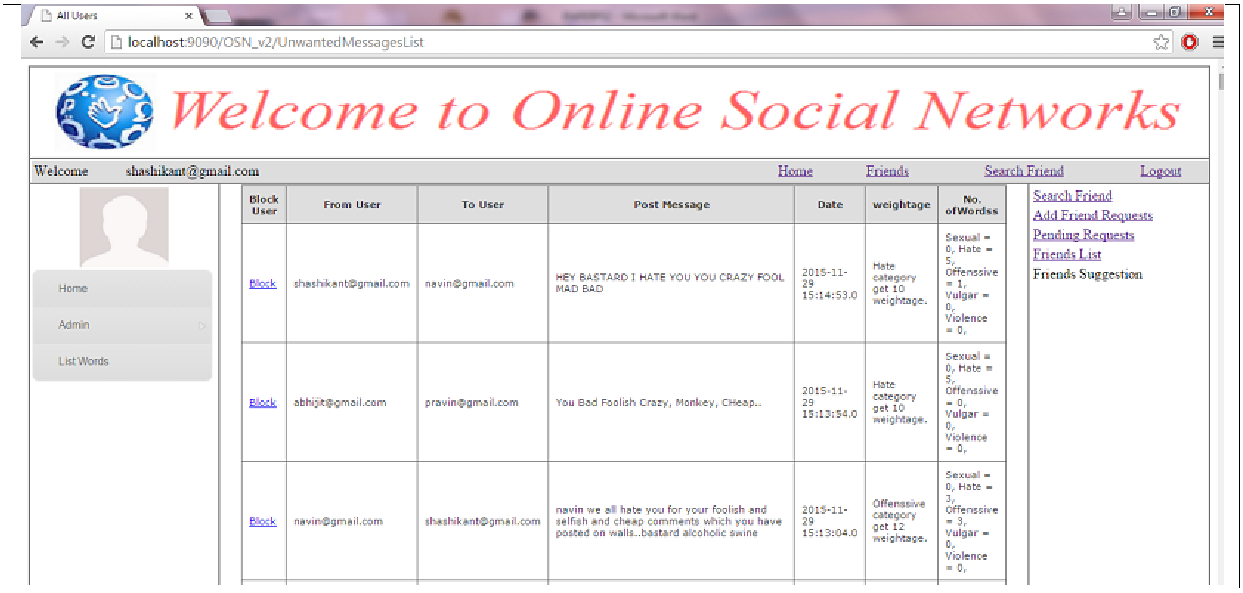
Vulgar

Enter Words:

Enter Words: 0

Id	Words	Category	wight	Delete
33	Fool	Hate	2	delete
34	Monkey	Hate	2	delete
35	Donkey	Hate	2	delete
36	bastard	Offenssive	4	delete
37	rascal	Offenssive	4	delete
38	Kill	Violence	6	delete
39	kidnap	Violence	6	delete
40	Stab	Violence	6	delete

Figure 4: Database of words categorize according category



Block User	From User	To User	Post Message	Date	weightage	No. ofWordss
Block	shashikant@gmail.com	navin@gmail.com	HEY BASTARD I HATE YOU YOU CRAZY FOOL MAD BAD	2015-11-29 15:14:53.0	Hate category get 10 weightage.	Sexual = 0, Hate = 5, Offenssive = 1, Vulgar = 0, Violence = 0,
Block	abhijit@gmail.com	pravin@gmail.com	You Bad Foolish Crazy, Monkey, CHeap..	2015-11-29 15:13:54.0	Hate category get 10 weightage.	Sexual = 0, Hate = 5, Offenssive = 0, Vulgar = 0, Violence = 0,
Block	navin@gmail.com	shashikant@gmail.com	navin we all hate you for your foolish and selfish and cheap comments which you have posted on walls..bastard alcoholic swine	2015-11-29 15:13:04.0	Offenssive category get 12 weightage.	Sexual = 0, Hate = 3, Offenssive = 3, Vulgar = 0, Violence = 0,

Figure 5: Unwanted Posts List

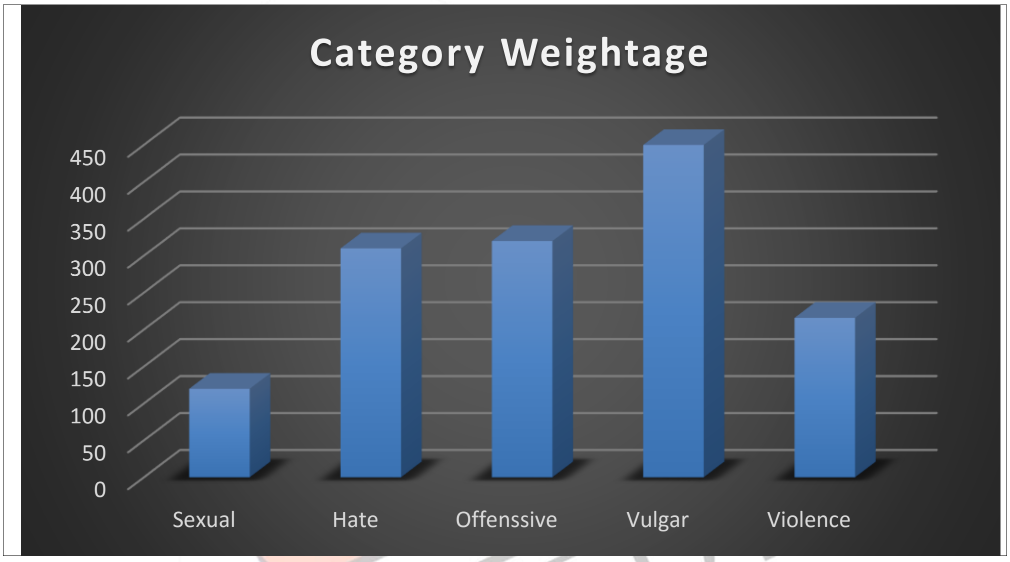


Figure 6: Shows graph of unwanted posts category to the weightage of posts.

VII. Conclusion

The systems approach is developed to filter unwanted messages from online social networks. Using classification and rule based approach undesired messages are recognized and discarded which further the decision is made that user should be inserted into blacklist. Filtering rules applied on the messages gives results more accurately and efficiently to display results. GUI of system provides handling easy to each and every user to understanding. This way we provide more privacy concern to user mentality and avoid such type of vulgar languages preventing to be posted on wall.

References

[1] M Vanetti, Elena Ferrari, Moreno Carullo, Barbara Carminati, Elisabetta Binaghi, and Barbara Carminati, “A System to Filter Unwanted Messages from OSNs User Walls,” 2013.

[2] M.Chau and H.Chen, “ A Machine Learning Approach to Web Page Filtering Using Content and Structural Analysis”, Decision Support Systems(dss), vol.44, no.2, pp.482-494, 2008.

[3] A.Adomavicius and G.Tuzhilin, “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the -Art and Possible Extensions,”.

[4] Elissa F.Sebastiani, “Machine Learning Automated Text Categorization”, ACMComputing survey, vol.34, no.1, pp.1-47, 2002.

[5] B.Sriram, D.Fuhry, M.Demirbas, and E.Demir, “Short Text Classification in Twitter to Improve Information Filtering”,.Proc.33rd Int’l ACM Conf. Research and Development in Information Retrieval(sIGIR ’10), 2010.

[6] V.Bobicev and M.Sokolova, “An Effective and Robust Method for Short Text Classification”, Proc.23rd National Conff. of Artificial Intelligence (AAAI), C.P. Gomes and D. Fox , eds., 2008.

[7] M.Vanetti, M.Carullo, E.Binaghi, and B.Carminati, “Content Based Filtering in On-Line Social-Networks”, 2010.

[8] J.Colbeck, “Combining Provenance with Trust in Social Networks for Semantic Web Content-Based Filtering”,,Proc. Int’l conf. Provenance and Annotation of Data, I. Fosters andL. Morea, eds., 2006.

[9] M.Carullo, E.Binaghi, and ,I.Gallo, “An Online Document Clustering Technique for Short Text Web Contents” Pattern Recognition Letter, vol.30, July 2009.

[10] C.D. Manning, H. Schu tze and P. Raghavan,“ Introduction to Information Retrieval(IR)”, Cambridge Uni.. Press, 2008.

[11] H.Schutze, J.O.Pedersen, and D.A.Hull,“ A Comparison of Classifiers and Document Representations for the Routing Problem ”,1995.

[12] R.J.Mooney and L.Roy, “Content-Based Book Recommending Using Learning for Text Categorizations”, 2000.

[13] S.Zelikovitz and H.Hirsh, “Improving Short Text Classification Using Unlabeled Background Knowledge”, Proc. 17th(ICML ’00) Int’l Conf. Machine Learning , P. Langley, ed., 2000.

[14] J.Moody and C. Darken, “Fast learning in Networks of Locally Tuned Processing Units”, neural Computation, vol. 1,no. 2,pp. 1989