Implementation of Content based Filtering of Undesirable Messages & Blocking Mechanism for OSN users

Ms. Komal A. Nazirkar
ME Student, Dept. of IT
SKNCOE, University of Pune
Pune, India

Prof. S. A. Nagtilak
Asst. Professor, Dept. of IT
SKNCOE, University of Pune
Pune, India

Abstract— **Online social networking has much more importance in daily life. User use OSN's to upload their information which may be public or private. Due to OSN's world becomes close to each other. The data uploaded by the user may contain some words which are abusive in nature or they might show some political or religious view. Due to all these things, the environment gets disturbed. To avoid all these things, a system is proposed which filter out the unwanted messages posted on their private wall called user wall. For that filtering rules and trust relationship criteria's are considered. Using both these criteria's user becomes able to restrict the messages posted on their wall. Also according to their behavior they get blocked temporarily or permanently.**

Keywords—Abusive, Blocking, Filtering, Online Social Networking, User Wall.

I. INTRODUCTION

Now days, people become close to each other due to OSN. OSN is online social networking. Over the past few years it becomes a very fast, useful and efficient way to remain in contact with each other. By using OSN, user can share a lot of information related to their personal lives. They can also share their information related to their relationship with each other. The shared information may contain some audio or video clips. Also the information may contain some data which may be public or private related to any specific user. If we considered the facebook in case of an OSN then while posting the messages on other user walls, may contain some which are abusive in nature, or they might show some political, or violence view. Because of this the environment gets disturbed.

In order to avoid all these things, we have proposed a system which filters out these undesirable messages. Using this system, user becomes able to restrict the unwanted messages which are going to post on their private wall called user wall. For that different filtering rules and trust relationship criteria are taken into consideration.

In this system, whenever user posts the message on other user walls, that message gets classified into different classes using text classifier algorithm (Naïve Bayes Algorithm explained in section IV). The main efforts are taken in building Short Text Classifier (STC) which is used for selection and extraction of data from the posted message. After that, filtering rules and trust relationship criteria are applied on that message. Filtering rules are the rules which are totally set by user and are set differently by different users [2]. While, trust relationship criteria means trust between different users [1]. If the facebook is considered then friends, friends of friends are come under trust relationship criteria. A threshold is calculated and according to that if the message is normal then only it gets posted on user wall otherwise gets filtered out by the system. A blacklist (BL's) is maintained to block the misbehaving users temporary or permanently. In case of temporary blacklisting, the users become unable to post on other user walls for some specific time period. While in case of permanent blacklisting, the user gets blocked permanently by the system [4].

II. RELATED WORKS

This section includes the survey which is done related to this system. There three methods for text filtering mainly collaborative filtering, content based filtering, and policy based personalization of OSN content. As the name suggested, the content of the message is filtered according to its content [3, 5]. In order to develop this system, there is a need of machine learning algorithm. There are lot of machine learning algorithms are available like Support Vector Machins (SVM's), Naive Bayes Algorithm, Decision trees. The performance of all these algorithms are calculated on the basis of real-time classification speed, learning speed, classification accuracy etc [7, 9].

The boosting algorithms can also be used for categorization of text. Boosting means to combine many simple and moderately inaccurate categorization rules into a single and highly accurate rule. There are different boosting algorithms for multi-label, multi class problems as

AdaBoost.MR and AdaBoost.MH. The AdaBoost.MR is basically designed to find a hypothesis which ranks the labels in a manner that hopefully places the correct labels at the top of the ranking while AdaBoost.MH algorithm is derived using a natural reduction of the multiclass, multi-label data to binary data [6].

A tool called Weka is used for data mining task which includes different types of machine learning algorithms. Weka contains tools for regression, data pre-processing, association rules, classification, clustering, and visualization [10].

III. PROPOSED SYSTEM

The main purpose of this proposed work is to develop a system which filters out undesirable messages from user wall in case of Online Social Networking. This is done using different algorithms. The architecture of the system is shown in fig.1.

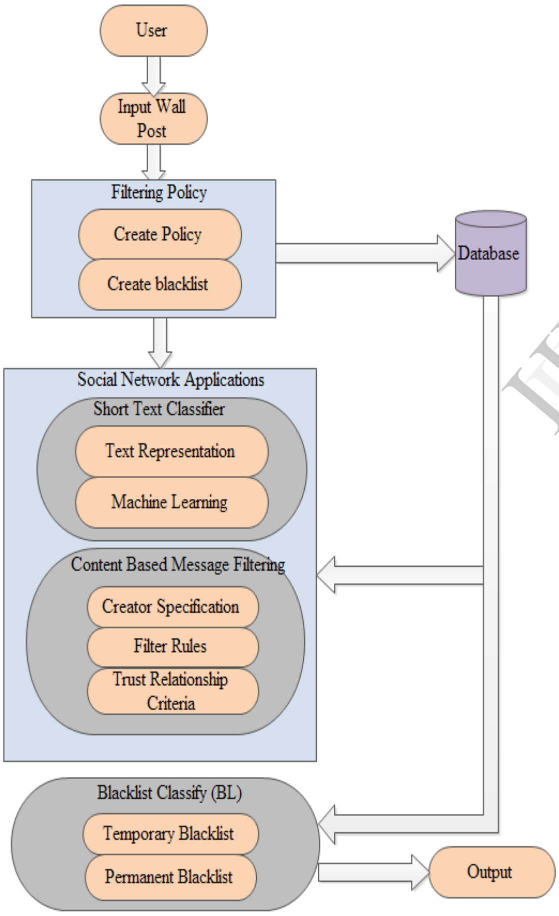

Fig.1 Proposed System

First of all user will enter the other user's wall and post the message. After that, short text classifier divides that message and classify into predefined classes. The filtering rules and trust relationship criteria are applied on that data. By using both these criteria's it is found that the posted message is normal then only it get posted on other user wall otherwise get filtered. If the is process going on then by considering trust relationship criteria, user get temporarily blacklisted. Then total friends of blacklisted user are calculated and that user gets permanently blocked if more than half of friends blacklisted him/her temporarily. The module wise description for this project is explained in next section.

IV. IMPLEMENTATION

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. Module 1: Text classification of posted message:

Whenever user posts the message on other user's wall that time, the message gets classified into different predefined classes Politics, Sports, Vulgar, Bollywood etc. It consists of four different stages in which first three stages are called pre-processing phase because they clean the input to make it ready for accurate analysis. If uncleaned input is passed to core algorithm (Stage 4) then results would be inaccurate**.** All these stages are explained as follows:

1) Stage 1: Stop word removal: This algorithm removes all the stop words (like a, an, the [articles], prepositions [on, at, etc.]) from the input text. The stop words do not convey any important information about the category of the message. Their removal ensures enrichment of input.

2) Stage 2: Stemming: Stemming means to reduce the words in the filtered text in stage (1) to their roots. E.g. "searching" is reduced to "search", "boys" is reduced to "boy", "advancement" to "advance". This algorithm enriches the input further making analysis more error resilient and accurate.

3) Stage 3: Case transformation: This transformation converts the text to the same case (either lower or upper) so that any text comparisons on word roots in stage (2). It doesn't fail due to case difference.

4) Stage 4: Categorization of text: This stage includes the core algorithm that categorizes input text in one of the previous trained categories. The preprocessed text is passed to the trained model for analysis. In the current version, stage 4 is primarily Naive Bayes classifier. Naïve bayes classifier is a probabilistic multiclass classifier [8]. The flowchart for naïve bayes classifier is shown in fig. 2.

In the given flowchart, first of all a cross validation method is calculated i.e. K-fold Cross Validation or holdout validation. After that input file is selected and that input file is read. If the success got then check if the cross Validation is

K-fold otherwise throw the error message. If the cross validation is K-fold then create training set and test set otherwise divide data set into K sets. After creation of training set and test set, the prior probability is calculated for training set. Prior probability is calculated if the description of the object is not known. Next step is the calculation of conditional probability for feature values in test data. After that, calculate posterior probabilities for each class. Posterior probability is calculated when description of object is known. Next step is to classification of the text and display of classification result. Then check that is the end of test set if yes then there is end of algorithm.

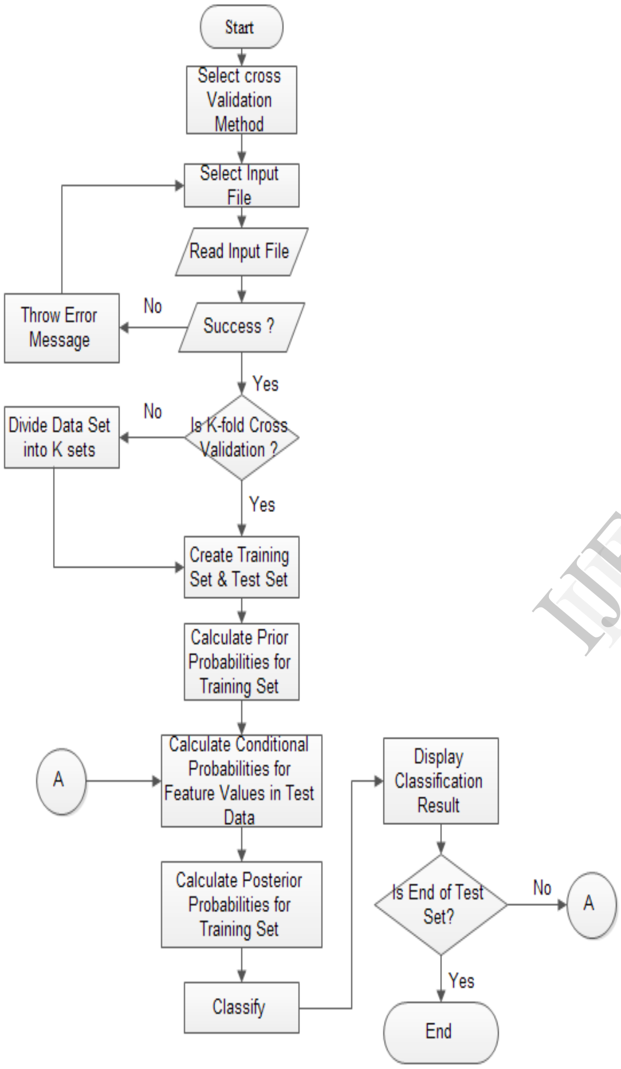
Start
Select cross Validation Method
Select Input File
Read Input File
Success ?
No
Throw Error Message
Yes
Is K-fold Cross Validation ?
No
Divide Data Set into K sets
Yes
Create Training Set & Test Set
Calculate Prior Probabilities for Training Set
A
Calculate Conditional Probabilities for Feature Values in Test Data
Calculate Posterior Probabilities for Training Set
Classify
Display Classification Result
Is End of Test Set?
No
A
Yes
End

Fig. 2 Flowchart for Naïve Bayes Algorithm

B. Module 2: Filtering of classified message:

After classification of posted message in different classes, the two criteria's are applied as:

1) User rules: These are the rules set by user itself. These rules may be different for different users. The user may change the rules from user to user. e.g. if there are three users like Alice, Bob and Demon then the filter rules set by Alice for Bob may be different for that of Demon.

2) Trust relationship: This is the next criteria, which is taken into consideration while filtering the message. This criterion is used to find the trust between two users. Because, it is obvious that trust between two friends is much more than that of a user and its friends of friend. So in order to find trust between two users several factors are considered which are explained as follows:

- GT: Global threshold
- RH: Recent history indicates number of days which will make "recent history" for checking recent offenses
- RO: number of recent offenses by sender (i.e. offenses committed by sender recently (in last RH days))
- FR: messages posted to the receiver which were filtered (all time)
- SFR: messages posted by a particular sender to this receiver which were filtered (all time)
- DTB: number of Days To Block after user crosses threshold.

The used threshold (TH) is calculated as follows (in sequence):

If the sender is not a friend of the receiver, global threshold is halved.

Thus, TH=GT/2

If offenses were committed by the sender recently, reduce TH by ratio of RO and RH. Thus,

$$TH=TH-(RO/RH)$$

If this sender has already offended the receiver in the past more than once, reduce the half of TH by FR/SFR ratio. Thus,

$$TH=TH-TH*FR/ (2*SFR)$$

Using these formulas, a threshold is calculated and according to that if user posts the normal message then only it gets posted on other user wall otherwise it gets filtered.

C. Module 3: Temporary blocking of misbehaving user:

In this module, the temporary blocking of misbehaving user is done. As shown in module 2, the threshold is calculated and according to that the no. of days to block misbehaving user is calculated which is the ratio of past offences and the threshold.

D. *Module 4: Permanent blocking of misbehaving user:*

After temporary blocking of misbehaving user the count for total no. of friends of that misbehaving user are calculated and according to that it is checked, how many friends have temporary blocked the same user.

Suppose, Bob is a misbehaving user and he has n1, n2, n3… friends. Hence, If no. of friends blocked Bob (temporary) > = n1+n2+n3…./2. Then, Bob will automatically get blocked permanently by the system. Because of this facility, Bob will not get access to his own account.

V. RESULTS AND DISCUSSIONS

The following Bar graph shows the accuracy comparison between existing system & proposed system.

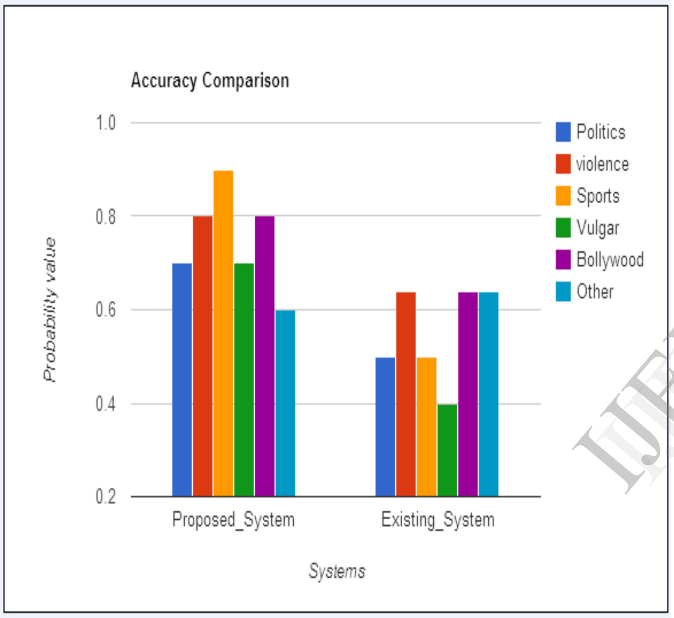

Fig. 3 Accuracy Comparison between Proposed & Existing System

A sample dataset is provided to each system & according to that accuracy is calculated. For that the pre-defined classes are considered on X-axis and the Probability value considered on Y-axis. Thus, from above graph it is clear that a better content based filtering system can be developed.

VI. CONCLUSION

From this project, it is concluded that a better system can be implemented in case of OSN. Using this system, it is possible to control unwanted messages posted on other user walls. This can be done using different filtering rules and trust relationship criteria. Whenever user posts the message on other user walls at that time if that message satisfies all the criteria then only it get posted on other user walls otherwise it get filtered by the system itself. It is also possible to block the misbehaving user temporarily so that he becomes unable to post on other user walls for a specific time period. If still the misbehaving user not behaves properly then it may be the possibility that he/she get permanently blocked by the system itself if more than half of the friends of misbehaving user blacklist the same user temporarily.

ACKNOWLEDGMENT

I am very thankful to my guide for guiding me and heartily thankful to IJARCSSE for giving me such a wonderful opportunity for publishing my paper.

REFERENCES

[1] J. Golbeck, “Combining Provenance with Trust in Social Networks for Semantic Web Content Filtering,” Proc. Int’l Conf. Provenance and Annotation of Data, L. Moreau and I. Foster, eds., pp. 101-108,2006.

[2] M. Vanetti, E. Binaghi, E. Ferrari, B. Carminati, and M. Carullo,, “A System to Filter Unwanted Messages from OSN User Walls”, IEEE Transactions on Knowledge and data Engineering, Vol. 25, No. 2, February 2013.

[3] M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari, “Content-Based Filtering in On-Line Social Networks, Proc. ECML/PKDD Workshop Privacy and Security Issues in Data Mining and Machine Learning (PSDML ’10), 2010.

[4] P. Tsang, A. Kapadia, “Nymble: Blocking Misbehaving Users in Anonymizing Networks”, IEEE Transaction on Dependable and Secure Computing, Vol. 8, No. 2, March-April, 2011.

[5] R.J. Mooney and L. Roy, “Content-Based Book Recommending Using Learning for Text Categorization”, Proc. Fifth ACM Conf. Digital Libraries, pp. 195-204, 2000.

[6] R. Schapire and Y. Singer, “Boostexter: A Boosting-Based System for Text Categorization”, Machine Learning, vol. 39, nos. 2/3, pp. 135-168, 2000.

[7] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, “Inductive Learning Algorithms and Representations for Text Categorization,” Proc. Seventh Int’l Conf. Information and Knowledge Management (CIKM ’98), pp. 148-155, 1998.

[8] S. Kotsiantis, V. Tampakas, “Increasing the accuracy of Hidden Naïve Byes model”, Advanced Information Management and Service (IMS), 6th International Conference, IEEE, pg. 247-252, 2010

[9] T. Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features”, Proc. European Conf. Machine Learning, pp. 137-142, 1998.

[10] http://www.cs.waikato.ac.nz/ml/weka/