Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 50 (2015) 466 – 475

Procedia Computer Science

2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

Automated Message Filtering System in Online Social Network

Subramaniyaswamy V[a], Logesh R[b],Vijayakumar V[c], and Indragandhi V[d]

[a,b]*School of Computing, SASTRA University, Thanjavur-613401, India*
[c]*School of Computer Science and Engineering, VIT University, Chennai-600127,India*
[d]*School of Electrical and Electronics Engineering, SASTRA University, Thanjavur-613401, India*

Abstract

In this generation, using online social network (OSN) is an unavoidable powerful weapon to exhibit peoples' views and ideas. The users depending upon their interests can select the persons who must post/comment messages in their wall. The present excavation in OSN user wall is "No filtering of abusive messages". That is the selected persons can post any sort of messages in their wall. So in this paper, we propose a filtered wall to permeate offensive messages using rule based and text classification techniques. We have evaluated the performance using metrics, from which it is shown that proposed method is better.

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

Keywords: OSN (Online Social Network); Expert analysis; Text classification; Text mining

1. INTRODUCTION

Social networking is basically a social interaction which relies upon personal relationships. It is a remarkable advancement on web which is highly needed for today's expeditious evolution. Social networking websites for instance Facebook, LinkedIn, and MySpace extends to us bountiful features. Let's discuss them below:

- Flashing idea of sharing images, videos, and files.
- Posting messages on friend's wall.
- Chat box that offers secrecy between two people.
- Conference chats with numerous people at a time.
- Following certain eminent persons or pages.
- Secure sharing of personal information with a definitive group of friends.

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)
doi:10.1016/j.procs.2015.04.016

- Imparting our likes and dislikes.

Social network holds an architecture that must have a profile, friends, groups, Discussions, Widgets. However all these are optional. A worthy point to note is that Social networks may also have a few detriments [Vanetti et al. 2013]. It becomes an open ground for hackers to commit faults and launch virus attacks. It may also result in scams and false use of data or information [Vanetti et al. 2011]. Sharing of information may sometimes be vulgar and it is inevitable. In this paper, we propose an idea of "Filtered wall" on the basis of personalized filtering recommendation algorithm [Weifeng et al 2011]. This notion is fundamentally developed to preclude the appearance abusive words on the user wall [Adomavicius and Tuzhilin 2005]. We adopt the concept of expert analysis wherein a third part is given the utmost importance. The third party serves as a critic and he checks with the available documents for the presence of those abusive words. If present, the word is blocked as per the third party's decision.

The persisting paper is devised as follows: Section 2 indulges in Features of OSN, Section 3 introduces the Filtered Wall, Section 4 describes the filtering rules and mechanisms, Section 5 illustrates the System Architecture, Section 6 demonstrates the mathematical proof for Machine Learning algorithm and Section 7 handles experimental results. At last, Section 8 reaches the conclusion of this paper.

2. RELATED WORK

The existing OSN has a marvelous characteristic by which users can customize their wall by restricting some of their friends to comment/post on it. Moreover, it also provides another feature called "Block List". Block List contains blocked users. If the user doesn't want to share their ideas to his/her particular friend (or) if the user doesn't wish to disclose their prevalence/details to their specific friends, then they can block permanently by this feature [Chau and Chen 2008], [Mooney and Roy 2000]. After they are blocked there will be no more relationship with each other in OSN.

Achieving customization in user wall, by restricting some users has lead to the growth of a new thorn called "Vulgarity of messages" [Sebastiani 2002]. To illustrate this issue, two years back, two college students from other state got arrested by the police since one student posted a scurrilous message about a reputed political party on another student's wall. This news disseminated over OSN like a viral disease and many people commenced commenting on the matter with various unpleasing words [Vanetti et al. 2010]. Meanwhile, the government performed eavesdropping to this matter and took ridiculous action against the students who were the initiative and who made this as a sensational issue. At present, to avoid this kind of problem, users need to block such kind of friends. But those friends may be user's relatives, close friends, well-wishers etc, who will also post harmless messages many times [Strater and Richter et al. 2007]. But no other solution existed in OSN to avoid this kind of blocking.

3. MESSAGE FILTERING SYSTEM

Information Filtering is one of the best solutions for the aforementioned issue. It can be implemented in OSN for a variety of reasons like in posting, commenting messages on walls. There exist numerous techniques to visualize information filtering for the removal of offensive/vulgar messages.

3.1. State-of-the-art research on Message Filtering

The main contribution of the proposed method is the design of the system which provides customizable content-based message filtering for the Online Social Networks, based on the Machine Learning techniques. Our work is related to the policy-based personalization for the OSN.

Content-Based Filtering

Message filtering systems is usually designed to classify the data and dynamically generated information dispatched

from the sender (i.e) information producer to the user and the received information should likely satisfy the requirements of him/her [Schapire and Singer 2000].
In the content-based filtering method, every user is assumed as independent. So, a content based filtering system selects the data based on the correlation between the user preferences and the content. This is completely opposite to the collaborative filtering system, which selects the items based on the correlation between the people with the similar user preferences.

For the most of earlier research, electronic mail has been preferred as the domain for the information filtering. Documents which were processed by the content-based filtering is mostly in the form of text [Nin et al. 2009]. This helps content-based filtering to come closer to the text classification. The filtering process can be designed as splitting the incoming document or information into relevant and non relevant. The complex filtering systems categorize the messages automatically as thematic categories [Zelikovitz and Hirsh 2000].

Content-based filtering is based on the Machine Learning techniques. By using ML technique, a classifier is automatically induced by learning from the pre defined examples. Recently, there is large number of related work has been available. These available methods usually differ in the feature extraction methods [Bobicey and Sokolova 2008]. Sometimes, it also differs in model learning and collection of samples too. Several experiments prove that Bag-of-Words approaches yield good performances. By considering the learning model, the major approaches are content-based filtering and text classification. These methods usually show similar advantages and disadvantages. Most of the text filtering method by the ML has been applied to the long-form text. The performance of the text classification methods depends on the nature of the text document [Sriram et al. 2010]. Applying the content-based filtering to the messages posted on the OSN user wall has challenges due to the short length of the messages. This short text classification has got high attention in the research community [Golbeck et al. 2006]. Recent work has spread lights on robustness, misspellings, nonstandard terms and noise [Uszok et al. 2004].

To overcome the disadvantage of Block List we provide "Black List" where users are blocked temporarily for few days. This feature gets invoked when the person continuously posts (more than 3 times) offensive messages on user's wall. Person in Black List cannot post messages on user wall for those temporary days. After that he/she will be able to post on the wall. **Fig.1** depicts these features. At any point of time, their relationship with user will not be disturbed. Before enclosing the person into Black List, profile checking will be done by taking into account their relationship strength (%), trust (%), kind of relationship (friend, family, etc) with the user.

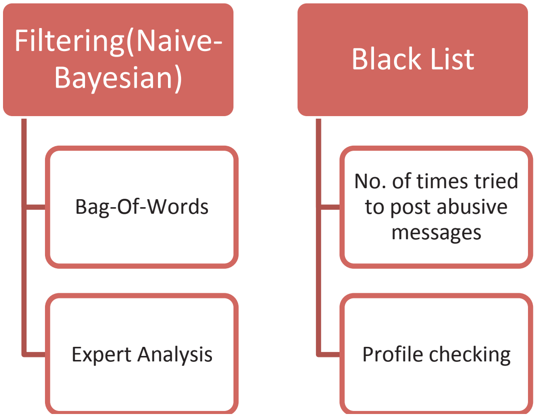
Filtering(Naive-Bayesian)
Bag-Of-Words
Expert Analysis
Black List
No. of times tried to post abusive messages
Profile checking

Fig.1: Filtering Mechanism and Black List

3.2. User registration and Login

In order to be a part of the Online Social Network, an user has to first get himself/herself registered in the Online social network. The registration process would consists of the user giving his/ her details like his name, age, occupation, mail id , so on which are meant to be genuine. The details that the user has registered in the Online Social Network are acquired and stored in the Data Store Blob Store, a GUI based Data Base that can store terabytes of data. The authenticated user will be permitted to access his home page once his user credentials are verified with the data store.

3.3.Friend request and Message posting on walls

The authenticated user on the Online Social Network has the privilege to send friend request to another authenticated user in the OSN by issuing a friend request to that particular person. Once the other person has accepted the friend request, then, both the users can share with each other their ideas in the form of pictures, status, videos. For each person in the friend list, user will maintain some of their details like relationship strength(%), trust(%), kind of relationship, etc..in a database. These details will be useful in case of “Black List” management.

4. FILTERING MECHANISMS AND RULES

Filtering is a process of removing unwanted material. Likewise in OSN the filtering is used to remove the unwanted messages, comments from being posted on the user’s wall [Kagal et al. 2004]. In our OSN, the people who are featured in the “Blocked List” cannot post on the user’s wall. Technically, the OSN uses the Naive Bayesian and rule based text classification along with “Bag-Of-words” (document with enormous words). Here, collection of offensive, abusive and vulgar words is framed as document through which first stage of filtering gets completed by checking the tokenized words (separated words that are taken from posted messages) against the document. If it is not present in the document, then we go for “Analysis by experts” (Third party consultation).

4.1.Filtering Rules

In the proposed system, filtering of abusive messages is done at the rule layer. This rule layer comprises of filtering rules. The proposed social network is modeled as a directed graph, in which each node is a network user and edges denotes the relationships between two different users. Edges were labeled by the type of relationship between the users such as friend, colleague, and parent. In addition trustworthiness is also labeled along with the relationship in the edges. To denote the trust levels rational numbers in the range {0,1} is used. So there will be two labels per edge. One is Relationship Type (RT) and another is trust value(X) between two users. If two users have indirect relationship, if there is a path of more than one edge connecting them, such type of edges have RT label alone. There are many algorithms already available can be used in our proposed system, since we do not address the problem of trust computation. Available algorithms mainly differ in the single criteria (i.e) to choose the paths on which trust computation should be done between the two users.

OSN message is represented as a characteristic vector adopting vector space model. The number of keywords in a message is n. First we extract the keywords which represented by w_i ($1 \leq i \leq n$) and the message text is denoted by t, where t=(w_1, w_2, ... , w_n). We use Naive Bayes algorithm to filter the OSN wall messages.

Naive Bayes algorithm comprises of two parts. They are, Bayesian training algorithm and Bayesian decision algorithm.
In the first part, the following steps are done.
1. Collect a great number of words classified by experts previously to construct two different training sets. One is worst words set and other is good words set.
2. Calculate and store the frequency of each word in the worst words set and good words set. These categories are denoted by c.
3. The times that a word w_i exists in the category c is denoted by N_i and the total number of words is n. Then the probability of the word w_i in the c can be calculated using following formula.

$$P(w_i) = \frac{\mathrm{Ni}}{\sum_{i=1}^{n} \mathrm{Ni}}$$

4. If event A has happened and the probability of event B happening is known as conditional probability of event B in the given event A. And it is denoted as P(B|A). M denotes the OSN message and c_j(j=1,2). C_j indicates the good

words set and worst words set. Conditional probability $P(w_i|c_1)$ and $P(w_i|c_2)$ can be calculated using above step. Then using those conditional probability, posterior probability $P(c_2|w_i)$ text belongs to worst words class using the following formula.

$$P(c_2|w_i) = \frac{P(wi|c2)}{P(wi|c2) + P\ (wi|c1)}$$

Establish a hash table(hash_possible). This hash_possible is used to store the keyword w_i and its corresponding posterior probability $P(c_2|w_i)$.
We only use the words in the worst words set as keywords to classify messages in order to improve the performance of the Naive Bayes algorithm for the text classification problem.

In the second part, the following steps are taken part.

1. The OSN message is treated with the word segmentation algorithm to get all words in the message. Then we search hash_possible to find which of words available in the hash_possible. M message is represented by $M=(w_1, w_2, ..., w_n)$. Here M is the message and it has n attributes.

2. After getting the keyword w_i, We can easily get the posterior probability $P(c_2|w_i)$ from hash_possible which is already generated. then we create the probability $P(M|w_1, w_2, ..., w_n)$, Where M belongs to Worst words class through the compond probability formula.

$$P(M|w_1, w_2, ..., w_n) = \frac{\prod_{i=1}^{n} P(c2|wi)}{\prod_{i=1}^{n} P(c2|wi) + \prod_{i=1}^{n}(1-P(c2|wi))}$$

The threshold is δ [Zhang and Callan 2001], The message will be classified as abusive when $P\ (M|w_1, w_2, ..., w_n) > \delta$.

5. SYSTEM ARCHITECTURE

The user to get into the OSN (Online Social Network) has to first enroll his/her details. Those details will be accumulated in the database. Then he can login with the registered username and password. The details entered by the user then will be verified with the data in the database. If the username and password matches perfectly with that of the registered one, then the user will be directed to the home page of the OSN. Otherwise, it will report a warning message to enter a valid username and password. There are many exiting features in OSN [Bizer and Cyganiak 2009], [Lewis et al. 2004]. The user can make friends and express their ideas and views. The relationship between various persons (like friend, family, etc) and their trust percentage will also be stacked in the database. The present work focuses mainly on the working of filtered wall which is briefly depicted in the above figure. First if the user posts/comments any message then it will be channelized to the filtered wall. Then pops the functioning of two fabulous techniques namely “Text Mining” and “Filtering” [Carullo et al. 2009]. The turnout of text mining leads to the classification of neutral (harmless) and non-neutral (abusive/vulgar) messages. **Fig.2** explains the entire working of the system.

Then the non-neutral messages will be handled by various methods like expert analysis, filtering rules and by Naïve-Bayesian classifier. Filtering rules is comprised of antecedents (LHS) that specifies conditions for the rule and consequents (RHS) that specifies the action that has to be executed for that particular condition.

Eg : (Document=Contains word) ^ (Expert = Worst) → Non-neutral

Expert analysis has to do with the decision of the third party whether a particular word is neutral or non-neutral. Then the final filtering will be confined by the Naïve-Bayesian classifier.

6. MACHINE LEARNING

To perform supervised learning using any machine learning algorithm, that particular algorithm must be trained to do classification, prediction etc. In case of unsupervised learning the task will be clustering [Golbeck 2005]. Training signifies granting particular known inputs to the algorithm for making it ready to take decision like classification, prediction etc for unknown inputs. This kind of training acts as a foundation for many machine learning algorithms like SVM, Neural networks and Bayesian. The foremost step in all these algorithms is to split the inputs into training set and test set [Jain et al. 2000]. Training set involves subset of input along with predicted class labels. Test set involves unpredicted subset of inputs.

6.1.Bayes Theorem

It is primarily utilized for decision making and inferential statistics that handles probability inference. It makes use of former events to predict the upcoming events.

P (h/D) = (P (D/h) P (h)) / P (D)

Where,

P(h) : Prior probability of hypothesis h
P(D) : prior probability of Training data D
P(h/D) : probability of h given D
P(D/h) : probability of D given h

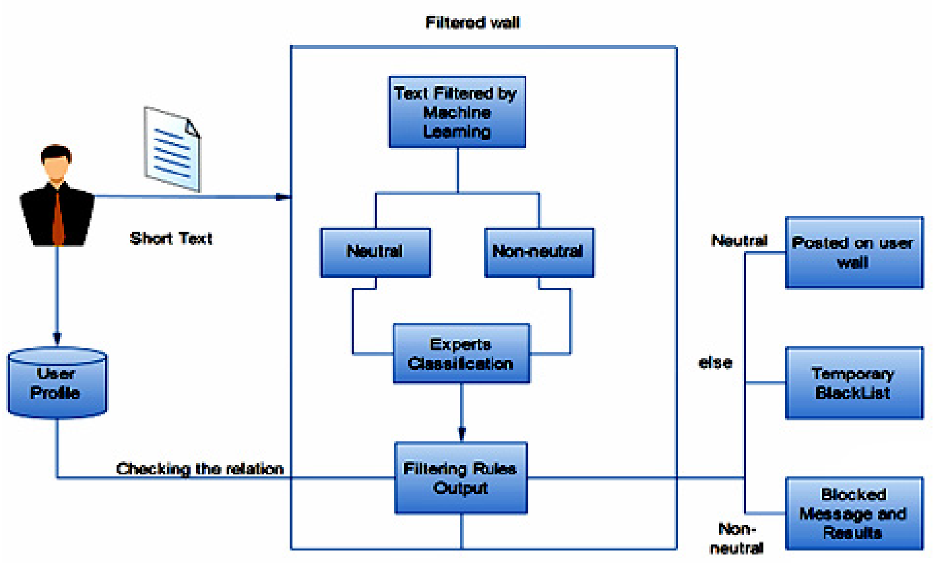

Fig.2: Overall Flow of the System

6.2.Maximum Likelihood (ML) Hypothesis Theorem

Suppose, assume the calculated probabilities are same, $P(h_i)=P(h_j)$
Then this may be reduced into : hML= arg max $P(D/h_i)$ Where, hi belongs to H.

6.3.Naive-Bayesian Classifier

It relies upon Bayesian theorem. It is well suited when the dimensionality of the inputs is high. Parameter estimation is done on the basis of maximum likelihood hypothesis [Hanani et al. 2001].

The superiority of Naïve-Bayesian is that:

- It requires only a pinch of training data to estimate the parameters.
- It is applicable even for complex real world problems.

6.4.Mathematical proof

Naïve Bayesian deals with categorical data so the posterior probability can be calculated by

$$P(a_i |v_j) = (nc+mp)/(m+n)$$

Where,
a_i – attribute values , i=0 to n
v- class label set
v_j – class membership values , j=1 to n
n – number of training examples for which v=vj
nc – number of examples for which v = vj & a=ai
p – priori estimate for p(aj | vj)
m – arbitrary constant

6.5.Pseudo code

Input Ts – Training Set with binary classlabel values ; acl – No. of record for which cl=cl_j ;
n – No.of training example for which cl =cl_j

Algorithm

```
Set ‘X’ to 3 [Arbitrary Constant]
Set ‘P’ to 0.5 [priori estimate for p( $a_i$ | $cl_j$ )
Get a[i], cl[j] where i,j are from 1 to n
For i 1 to n do
For j 1 to n do
        $P( a_i | cl_j) = (acl + xp) / (x + n)$
        End for
End for
Set $cl_j$ to “post” [class label value-1]
For i 1 to n do
        For j 1 to n do
              If $cl_j$ equals post then
                      Set $P( a(cl_j) )$ to $P( cl_j) *P( a_i | cl_j)$
              Else
                      Set $P( b(cl_j) )$ to $P( cl_j) *P(a_i | cl_j)$
              End If
        End for
End for
Set MLC to argmax $(P(a(cl_j)) , P(b(cl_j)) )$
```

Output MLC [Most likelyhood classification]

7. EXPERIMENTAL RESULTS

The filtering process in the proposed system has been exercised involving Naïve-Bayesian algorithm, filtering rules and expert analysis. The messages/comments posted in the user wall is pushed into the filtered wall. The filtered wall has numerous interesting features that include three experts for examining the incoming messages and options to include text files which contain some abusive words. The outlook of the filtered wall is depicted which plays a major role in this paper.

By choosing the experts it channelizes to select any one of the two options namely “view new comments” and “view blocked comments”. Selecting the menu “Import File” permits them to upload any text file containing bad words. If the incoming message matches with any word in that of the uploaded file then it will be displayed in “view blocked comments” menu describing the appropriate category of the text file in which it was present. The categorization of non-neutral messages combined with experts’ suggestions. Also to enhance the system the three experts after considering the category can further submit their percentage for allowance of the word as per their view. A threshold value will be maintained for making decisions preferred by the experts. All the percentages given by the experts will be recorded in the data store and an average of all the percentage will be calculated and will be compared with the threshold value. If the computed value exceeds the threshold value then it will be blocked.

Whereas, if no matches occurred then the message will be found in “view new comments” menu which will have two options called allow and block which can be chosen by the administrators. If the experts feel that the particular message contains abusive messages then they can go for block option or else if they choose allow option the message will then be posted in the appropriate user wall. The majority among the decisions taken by the three experts will be taken into account.

7.1.Performance Evaluation

RBFN and Naïve Bayesian deals with both supervised & unsupervised learning. RBFN has three layer namely inner layer, middle hidden layer & outer layer. The middle layer comprises of ‘n’ no. of neurons which requires continuous values for making the prediction because the calculation revolves around Gaussian function. So, more precisely RBFN is suitable for continuous values rather than categorical values for this kind of online filtering. But our work focuses on categorical values which are captured perfectly by Naïve Bayesian. Moreover, it requires only small set of training set. We have constructed a training set with keywords of worst words set and good words set. A brief comparison is pictorially represented below in **Fig.3 and Fig.4.**

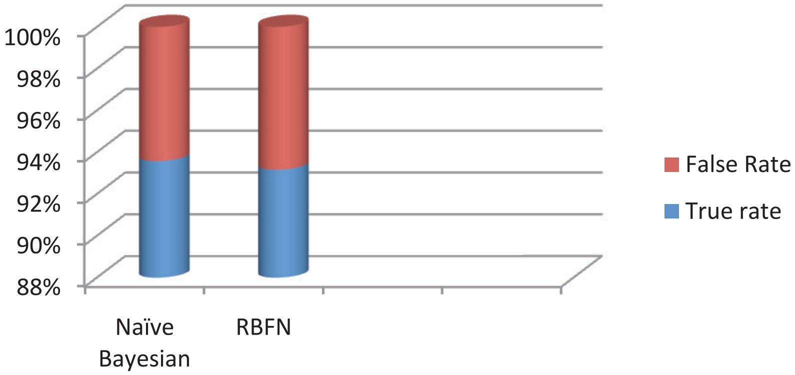

Fig.3: Overall Comparison between two algorithms

We have evaluated our method using two indexes. One is precision ratio R_c. R_c is calculated by,

$$R_c = \frac{Nr\_correct}{Nr\_total}$$

Where, Nr_correct indicates the number of abusive messages detected and Nr_total indicates the total number of abusive messages in the training set. The other is false alarm ratio R_W. R_W is calculated by,

$$R_w = \frac{\text{No\_wrong}}{\text{No\_total}}$$

Where, No_wrong indicates the number of normal messages in the training set which is classified wrongly and No_total indicates the total number of normal messages in the training set.

We have also used an evaluation metric which is most commonly used in the text classification problems. It is F-measure, which can be calculated using the following formula.

$$F = \frac{(\beta 2+1).Rc.(1-Rw)}{\beta 2.Rc+(1-Rw)}$$

Table 1 defines the statistical analysis of the proposed model and Table 2 is the quality metrics of the online social network message filtering system.

Table 1 : Statistical Analysis

Algorithm	Correctly classified instances %	Incorrectly classified instances %	Time taken (seconds)
Naive Bayesian	93.25	6.75	0.47
RBFN	92.85	7.15	0.63

Table 2: Quality Metrics

Metrics / Algorithm	RBFN	Naive Bayes Algorithm
Precision	98.755 %	97.78 %
False Alarm Ratio	25.95 %	16.58 %
F-Measure	74.49 %	83.72 %

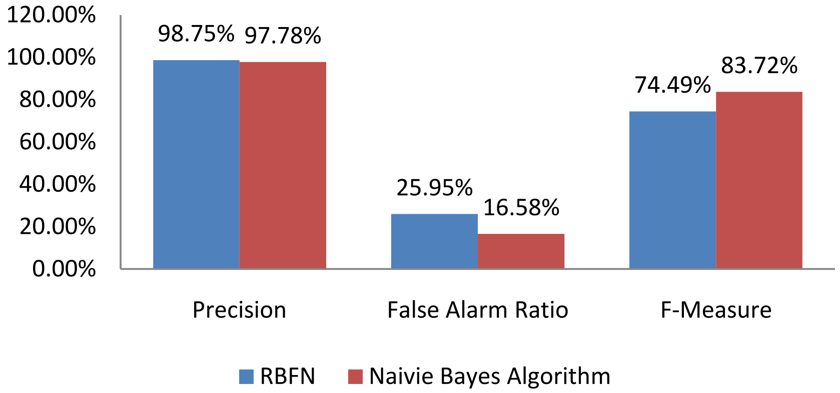

Fig.4: Comparison between two algorithms on basis of quality metrics

8. Conclusion

By this Proposed System, we enforce powerful techniques to achieve filtering of messages in OSN user walls. Naive-Bayesian is one of the best machine learning text classifier. It earned such popularity since it is well suited for online filtering such as spam mail filtering. This method is very easy to implement which reaps an excellent performance with very less complexity. Even though it is a very simple approach it provides an effective efficiency which cannot be satisfied by the existing algorithms. The future work concentrates on image and video filtering and further development can be done to change the learning process by making the machine itself to get trained by online learning mechanism which will replace the usage of database with terabytes of size since it is highly dynamic which cannot be weighed with small set of data.

References

1. M.Vanetti, E.Binaghi, B.Carminati, E.Ferrari and M.Carullo, "A System to Filter Unwanted Messages from OSN User Walls", IEEE Transactions on Knowledge and Data Engineering", **25**, 285-297 (2013).
2. M.Vanetti, E.Binaghi, B.Carminati, M.Carullo and E.Ferrari, "Content-Based Filtering in On-line Social Networks", Privacy and Security Issues in Data Mining and Machine Learning Lecture Notes in Computer Science, **6549**, 127-140 (2011).
3. S.Weifeng, S.Mingyang, L.Xidong and L.Mingchu, "An Improved Personalized Filtering Recommendation Algorithm", Applied Mathematics & Information Sciences, **5**, 69S-78S (2011).
4. A.Adomavicius and G.Tuzhilin, “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions,” IEEE Trans. Knowledge and Data Eng., **17**, 734-749 (2005).
5. M.Chau and H.Chen, “A Machine Learning Approach to Web Page Filtering Using Content and Structure Analysis,” Decision Support Systems, **44**, 482-494 (2008).
6. R.J.Mooney and L.Roy, “Content-Based Book Recommending Using Learning for Text Categorization,” Proc. Fifth ACM Conf. Digital Libraries, 195-204 (2000).
7. F.Sebastiani, “Machine Learning in Automated Text Categorization,” ACM Computing Surveys, **34**, 1-47 (2002).
8. 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).
9. K. Strater and H. Richter, “Examining Privacy and Disclosure in a Social Networking Community,” Proc. Third Symp. Usable Privacy and Security (SOUPS ’07), 157-158 (2007).
10. R.E.Schapire and Y.Singer, “Boostexter: A Boosting-Based System for Text Categorization,” Machine Learning, **39**, 135-168 (2000).
11. J.Nin, B.Carminati, E.Ferrari, and V.Torra, “Computing Reputation for Collaborative Private Networks,” Proc. 33rd Ann. IEEE Int’l Computer Software and Applications Conf., **1**, 246-253 (2009).
12. S.Zelikovitz and H.Hirsh, “Improving Short Text Classification Using Unlabeled Background Knowledge,” Proc. 17th Int’l Conf.Machine Learning (ICML ’00), P.Langley, ed., 1183-1190 (2000).
13. V.Bobicev and M.Sokolova, “An Effective and Robust Method for Short Text Classification,” Proc. 23rd Nat’l Conf. Artificial Intelligence (AAAI), D.Fox and C.P. Gomes, eds., 1444-1445 (2008).
14. B.Sriram, D.Fuhry, E.Demir, H.Ferhatosmanoglu, and M.Demirbas, “Short Text Classification in Twitter to Improve Information Filtering,” Proc. 33rd Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’10), 841-842 (2010).
15. 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., 101-108 (2006).
16. A.Uszok, J.M.Bradshaw, M.Johnson, R.Jeffers, A.Tate, J.Dalton, and S.Aitken, “Kaos Policy Management for Semantic Web Services,” IEEE Intelligent Systems, **19**, 32-41 (2004).
17. L.Kagal, M.Paolucci, N.Srinivasan, G.Denker, T.Finin, and K.Sycara, “Authorization and Privacy for Semantic Web Services,” IEEE Intelligent Systems, **19**, 50-56 (2004).
18. Y.Zhang and J.Callan, “Maximum Likelihood Estimation for Filtering Thresholds,” Proc. 24th Ann. Int’l ACM SIGIR Conf.Research and Development in Information Retrieval, 294-302 (2001).
19. C.Bizer and R.Cyganiak, “Quality-Driven Information Filtering Using the Wiqa Policy Framework,” Web Semantics: Science, Services and Agents on the World Wide Web, **7**, 1-10 (2009).
20. D.D.Lewis, Y.Yang, T.G.Rose, and F.Li, “Rcv1: A NewBenchmark Collection for Text Categorization Research,” J. Machine Learning Research, **5**, 361-397 (2004).
21. M.Carullo, E.Binaghi, and I.Gallo, “An Online Document Clustering Technique for Short Web Contents,” Pattern Recognition Letters, **30**, 870-876 (2009).
22. J.A.Golbeck, “Computing and Applying Trust in Web-Based Social Networks,” PhD dissertation, Graduate School of the Univ. of Maryland, College Park, (2005).
23. A.K.Jain, R.P.W.Duin, and J.Mao, “Statistical Pattern Recognition:A Review,” IEEE Trans. Pattern Analysis and Machine Intelligence, **22**, 4-37 (2000).
24. U.Hanani, B.Shapira, and P.Shoval, “Information Filtering: Overview of Issues, Research and Systems,” User Modeling and User-Adapted Interaction, **11**, 203-259 (2001).