International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Unwanted Message Filtering in OSN User Walls

Ashwini Swami[1], Jayashri Chaudhari[2]

[1]Pune University, BSIOTR (W), Wagholi, Pune, Maharashtra, India

Abstract: ***As we all know, On-line Social Networks (OSNs) are very popular medium to communicate, share and a considerable amount of human life information. Daily and continuous communications, But it is very important for On-line social networks, to avoid unwanted messages getting displayed onto user walls. But OSNs provide very little support to this requirement. To achieve this requirement, we propose a system letting OSN users to have a direct control on the messages posted on their walls. We did this through a flexible rule-based system, which allows users to adapt the filtering criteria to be applied to their walls, and a Machine Learning classifier technique which automatically generates membership labels in support of content-based filtering.***

Keywords: Information Filtering, On-line Social Networks, Short Text Classification, Filtering, Personalization, Machine learning classification techniques, Profile learning.

1. Introduction

As we all know, On-line Social Networks (OSNs) are best medium to communicate, and share information across the internet. Through this, one can exchange several types of content, like text, image, audio and video data. According to survey, in Facebook per month more than 30 billion pieces of content are shared. So, there is a need for user to avoid or to filter unwanted message or message containing bad words like vulgar words or abusing words. But OSNs provide very little support to this requirement. Just consider Facebook, in which users can only state who is permitted to insert messages in their walls like i.e., friends, friends of friends, or defined groups of friends. But no support for content-based preferences. Traditional classification techniques are not suitable for short text messages, as these short messages do not provide sufficient word occurrences.

Therefore here our purpose is to offer a system, which automatically filters unwanted messages from OSN user walls This is a flexible rule-based system, which allows users to adapt the filtering criteria to be applied to their walls, and a Machine Learning classifier technique which automatically generates membership labels in support of content-based filtering. We use Machine Learning (ML) text categorization techniques to allot each short text message to its respective category. In section 2 we describe surveys related work, 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 describes results. Finally, section 7 concludes the paper.

2. Literature Survey

Foltz and dumais studied different tested methods for determining which Technical Memos (TMs) best match people's technical interests. Within Bellcore, nearly 150 new TMs are published every month, yet very few are related to any single person's interests. Feedback using previously used abstracts provided an efficient and easy way of demonstrating people's interests [1]. There was a total focus on previous feedback. No individual based filtering was found there.

Filtering depends on explanations of individual or group information preferences that typically represent long-term interests. Users get only the data that is mined. Information filtering systems are aimed to classify a stream of dynamically generated information and present it to the user that information that is likely to satisfy user's requirements [2]. The paper was having main focus on the similarity between Information filtering and Information retrieval.

For microblogging services like twitter, to avoid overwhelming users by raw data a classification method has been offered to classify short text messages. One solution to solve this problem is to classify the short text messages [3]. The results obtained by proposed approach are not better.

The work by J. Golbeck presented an application, called FilmTrust, to personalize access to the website. But, there is no filtering policy layer for the result of the classification process to decide how and at which extent filtering out unwanted information. [4.] As far as privacy is concerned, current work is mainly focusing on privacy-preserving data mining skills, that is, protecting information related to the network. A user chooses a discrete rating value (e.g., not interesting, somewhat interesting, no comment, very interesting etc.) for each document read in the trial filtering system. A learning algorithm is used to associate these user ratings with document features and shared ratings from earlier readers to prioritize incoming information [5]. This paper is only concerned to diversified domains like Internet "news" articles, newswire articles, and broader network resources. This paper provided attention on just prioritizing information by making use of rating values.

The work by Boykin and Roychowdhury [6] that offered an automated anti-spam tool that can recognize unsolicited commercial e-mail, spam and messages related with people the user knows. However, it is important to note that the strategy just stated does not make use of ML content-based techniques.

3. Conceptual Architecture of Filtered Wall

The OSN architecture services are a three-tier structure (Figure 1). Social Network Manager (SNM) is the first layer, which serves to provide the basic OSN functionalities; the

second layer provides the support for external Social Network Applications (SNAs). For supported SNAs there is a requirement of an additional layer for their desired Graphical User Interfaces (GUIs). By considering this reference architecture our proposed system is placed in the second and third layers. To set up and manage FRs/BLs users interact with the system through GUI. The filtered wall is presented to the user through the GUI, where only authorized messages are published. There are two main components of our proposed system: Content-Based Messages Filtering (CBMF) and the Short Text Classifier (STC) modules. The function of STC is to classify messages according to a set of categories.

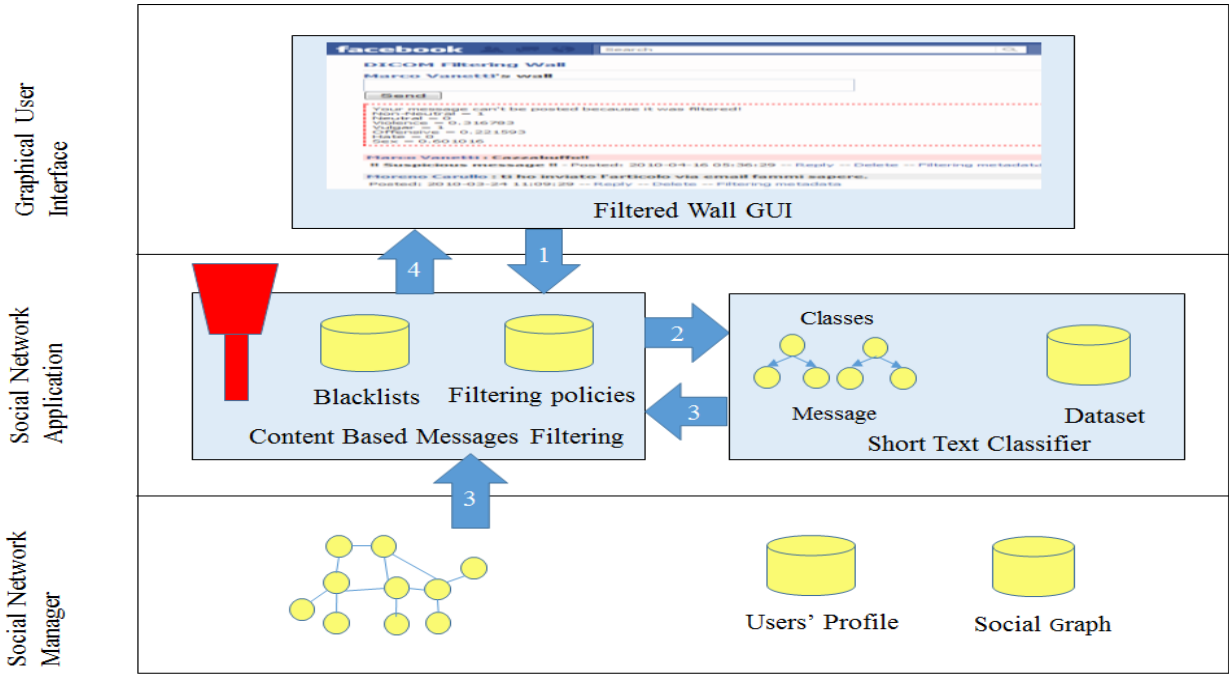

Figure 1: Conceptual Architecture of Filtered Wall

As shown in Figure 1, the overall architecture flows given as follows:

1) The user writes and tries to post a message after entering the private wall of his/her contacts which is interrupted by FW.
2) From this message content a ML based text classifier extracts metadata.
3) A metadata together with data extracted from the social graph and users' profiles provided by the classifier is used by FW, to impose the filtering and BL rules.
4) Based on the result of the previous step the message will be published or filtered by FW.

4. Short Text Classifier

The short text classifier we used is a hierarchical two level classifier which first identifies and eliminates "neutral" sentences, then classify "non neutral" sentences. The first level task is a hard classification where short texts are labeled neutral and Non-Neutral labels. The second level task is soft classification, which produces "gradual membership" for each of the appropriate classes, without taking any "hard" decision on any of them. Such a list of grades is then used by the successive phases of the filtering process.

4.1Text Representation

We consider three types of features, BoW, Document properties (Dp) and Contextual Features (CF). The BoW and Dp features are endogenous, and already used in [9]. Text representation using endogenous knowledge has a good general applicability, though in operational settings it is appropriate to use also exogenous knowledge. We present contextual features (CF) modelling information that depicts the environment where the user is posting. These features are very important in deterministically understanding the semantics of the messages [12].

According to Vector Space Model (VSM) for text representation, a text document d_j is denoted as a vector of binary or real weights $d_j = w_{1j}, \ldots, w_{|T|j}$, where T denotes set of terms that occur at least once in at least one document of the collection Tr, and $w_{kj} \in [0; 1]$ denotes how much term t_k contributes to the semantics of document d_j. In the BoW representation, terms are identified with words. For non-binary weighting, the weight w_{kj} of term t_k in document d_j is computed according to the standard term frequency - inverse document frequency (tf-idf) weighting function, defined as

$$tf - idf(t_k, d_j) = \#(t_k, d_j).\log\frac{|T_r|}{\#T_r(t_k)} \quad (1)$$

Where # (t_k, d_j) represents the number of times t_k occurs in d_j, and #T_r (t_k) represents the document frequency of term t_k, i.e., the number of documents in T_r in which t_k occurs. Dp features are heuristically calculated; their definition stems from intuitive considerations, domain specific criteria and in some cases required trial and error procedures.

4.2. Machine Learning-Based Text Classification

Short text classification is a hierarchical two-level classification process. The first-level classifier accomplishes a binary hard classification that labels messages as Neutral and Non-Neutral. In second-level task a finer-grained classification is performed. The second-level classifier accomplishes a soft-partition of Non-neutral messages assigning a given message a gradual membership to each of the non-neutral classes. With respect to other state of the art classifiers we select the RBFN model, among the variety of multi-class ML models well-suited for text classification. Let Ω be the set of classes to which each message can belong to. Each element of the supervised collected set of messages D = $\{(m_i, \vec{y}_i). . . (m_{|D|}, \vec{y}_{|D|})\}$ is composed of the text m_i and the supervised label $\vec{y}_i \in \{0,1\}^{|\Omega|}$ describing the belongingness to each of the defined classes. The set D is then into two partitions, namely the training set TrSD and the test set TeSD. Let M1 and M2 be the first and second level classifier, respectively, and $_i$ be the belongingness to the Neutral class. The learning and generalization phase works as follows:

1) we extract the vector of features $_i$ from each message m_i. The two sets TrSD and TeSD are then converted into TrS = $\{(\vec{x}_i, \vec{y}_i), . . . ,(\vec{x}_{|TrSD|}, \vec{y}_{|TrSD|})\}$ and TeS = $\{(\vec{x}_i, \vec{y}_i), . . . ,(\vec{x}_{|TeSD|}, \vec{y}_{|TeSD|})\}$, respectively.
2) for M1, a binary training set TrS1 = $\{(\vec{x}_j, \vec{y}_j) \in TrS \mid (\vec{x}_j, y_j), y_j = \vec{y}_{j1}\}$ is created.
3) for M2, a multi-class training set TrS2 = $\{(\vec{x}_j, y_j) \in TrS \mid (\vec{x}_j, \vec{y}'_j), \vec{y}'_{jk} = \vec{y}_{jk+1}, k=2, . . ., |\Omega|\}$ is created.
4) to distinguish whether or not a message is non-neutral, M1 is trained with TrS1. The performance of the model M1 is then calculated using the test set TeS1.
5) to calculate the gradual membership to non-neutral classes, M2 is trained with the non-neutral TrS2 messages. Then the performance of the model M2 is calculated using the test set TeS2. Therefore, the hierarchical system is composed of M1 and M2, where the overall computed function f: Rn → $R^{|\Omega|}$ is able to map the feature space to the class space, that is, to recognize the belongingness of a message to each of the $|\Omega|$ classes.

5. Filtering Rules and Blacklist

Here, we introduce the rules adopted for filtering unwanted messages. We model a social network as a directed graph. In this directed graph each node denotes a network user and edges denotes relationships between two different users. Each edge is labeled by the type of the established relationship (e.g., friend of, colleague of, parent of) and, possibly, the corresponding trust level, which represents how much a given user is trustworthy. We believe that trust levels are rational numbers in the range [0; 1]. There exists a direct relationship of a given type RT and trust value X between two users, if there is an edge connecting them having the labels RT and X. More than one edge connecting two users represents an indirect relationship.

5.1. Filtering Rules

The same message may have different meanings and depends on who writes it. Message creators on which a FR are to be applied can be selected on the basis of several different criteria, like imposing conditions on their profile's attributes.
Creator specification, defined as follows.

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

1) A set of attribute constraints of the form an OP av, where an is a user profile attribute name, av and OP are, respectively, a profile attribute value and a comparison operator, compatible with an's domain.
2) A set of relationship constraints of the form (m; rt; minDepth; maxTrust), indicating all the OSN users participating with user m in a relationship of type rt, having a depth greater than or equal to minDepth, and a trust value less than or equal to maxTrust.

Example - The creator specification CS1 = {Age < 17; Sex = male} denotes all the males whose age is less than 17 years, whereas the creator specification CS2 = {Dipika; colleague; 2; 0:4} denotes all the users who are colleagues of Dipika and whose trust level is less than or equal to 0.4. Finally, the creator specification CS3 = {(Dipika; colleague; 2; 0:4); (Sex = female)} selects only the female users from those identified by CS2.

For content-based filtering criteria, we make use of the two-level text classification to identify messages that, with high probability, are neutral or non-neutral.; as well as, in a similar way, messages dealing with a particular second level class. The last component of a FR is the action. The possible actions are "block" and "notify", with the obvious semantics of blocking the message, or notifying the wall owner and wait him/her decision. A FR is therefore formally defined as follows.

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

1) author is the user who shapes the rule;
2) creatorSpec is a creator specification, specified according to Definition 1;
3) contentSpec is a Boolean expression defined on content constraints of the form (C, ml), where C is a class of the first or second level and ml is the minimum membership level threshold necessary for class C to make the constraint satisfied;
4) action ∈ {block; notify} represents the action to be performed by the system on the messages matching contentSpec and created by users recognized by creatorSpec.

More than one filtering rule can apply to the same user. A message can be published only if it is not blocked by any of the filtering rules that apply to the message creator.

5.2. Online Setup Assistant for FRs Threshold Calculation

By Online Setup Assistant (OSA) procedure, we solve the problem of setting thresholds to filter rules. Messages selected from the dataset is presented to the user through OSA. For each message, the user states the system the willingness or the decision to accept or reject the message. The collection and processing of user decisions on sufficient set of messages distributed over all the classes permits to compute customized thresholds signifying the user attitude in accepting or rejecting certain contents.

Following process is used to select such messages. A particular amount of non-neutral messages taken from a fraction of the dataset and not belonging to the training/test sets. These messages are categorized by the ML in order to have, the second level class membership values. Then class membership values are quantized into a number of qC discrete sets. And, for each discrete set, we select a number nC of messages, obtaining sets MC of messages with |MC|= nCqC, where C € Ω - {Neutral} is a second level class. For instance, for the second level class Violence, we select 6 messages belonging to 8 degrees of Hate, for a total of 48 messages. For each second level class C, messages belonging to MC are shown. For each displayed message m, the user is asked to tell the decision ma € {Filter, Pass}. This decision indicates the readiness of the user to filter or not filter the message. Together with the decision ma the user is asked to express the degree of certainty mb € {0, 1, 2, 3, 4, 5} with which the decision is taken, where mb = 5 represents the highest certainty, and mb = 0 shows the lowest certainty.
Example 1. Suppose that Dipika is an OSN user and she wants to always block messages having a high degree of hate content. Through the session with OSA, the threshold representing the user attitude for the hate class is set to 0.7. Now suppose that Helen wants to filter only messages coming from indirect friends, whereas for direct friends such messages should be blocked only for those users whose trust value is below 0.3.

5.3. Blacklists

Blacklists (BLs) are directly managed by the system, which should be able to decide the users to be inserted in the BL and decide user's retention in the BL is finished. Such information is provided to the system through a set of rules, called BL rules. We let the wall's owners to state BL rules regulating who has to be banned from their walls and for how long. Therefore, a user might be banned from a wall, by, at the same time, being able to post in other walls.

Like FRs, our BL rules make the wall owner able to recognize users to be blocked based on their profiles as well as their relationships in the OSN. Therefore, by means of a BL rule, wall owners can be able to bar from their walls users they do not directly know BL for another while, as his/her behaviour is not improved. This principle works for those users that have been already inserted in the considered BL at least one time.

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

1) author is the OSN user who states the rule, i.e., the wall owner;
2) creatorSpec is a creator specification, specified according to Definition 1;
3) creatorBehavior consists of two components RFBlocked and minBanned. RFBlocked = (RF, mode, window) is defined such that:
   - RF = $\frac{\#bMessages}{\#tMessages}$, where #tMessages is the total number of messages that each OSN user recognized by creatorSpec has tried to publish in the author wall (mode = myWall) or in all the OSN walls (mode = SN); whereas #bMessages is the number of messages among those in #tMessages that have been blocked;
   - window is the time interval of creation of those messages that have to be considered for RF computation; minBanned = (min, mode, window), where min is the minimum number of times in the time interval specified in window that OSN users identified by creatorSpec have to be inserted into the BL due to BL rules specified by author wall (mode = myWall) or all OSN users (mode = SN) in order to fulfil the constraint.
4) T indicates the time period the users identified by creatorSpec and creatorBehavior have to be banned from author wall.

Example 2. The BL rule:
(Prabhu; (Age < 17); (0:5; myWall; 2 week); 2 days) inserts into the BL associated with Prabhu's wall those young users (i.e., with age less than 17) that in the last week have a relative frequency of blocked messages on Remo's wall greater than or equal to 0:5. Furthermore, the rule states that these banned users have to stay in the BL for two days. If Remo adds the following component (4,SN, 2 week) to the BL rule, he enlarges the set of banned users by inserting also the users that in the last week have been inserted at least four times into any OSN BL.

6. Results and Discussion

The analysis of related work has highlighted the lack of a publicly available benchmark for comparing different approaches to content based classification of OSN short texts. To cope with this lack, we have built a dataset D of messages. 1266 messages from publicly accessible groups have been selected and extracted by means of an automated procedure that removes undesired spam messages and, for each message, stores the message body and the name of the group from which it originates. The messages come from the group's web page section, where any registered user can post a new message or reply to messages already posted by other users. The set of classes considered in our experiments is Ω = {Neutral; Violence; Vulgar; Offensive; Hate; Sex}, where Ω – {Neutral} are the second level classes. The percentage of elements in D that belongs to the Neutral class is 31%. In order to deal with intrinsic ambiguity in assigning messages to classes, we conceive that a given message belongs to more than one class. Each message has been labeled by a group of five experts and the class membership values y_j € {0,1}Ω for a given message m_j were computed by a majority voting procedure. After the ground truth collection phase, the messages have been selected to balance

as much as possible second-level class occurrences. The group of experts has been chosen in an attempt to ensure high heterogeneity concerning sex, age, employment, education and religion. In order to create a consensus concerning the meaning of the Neutral class and general criteria in assigning multi-class membership we invited experts to participate to a dedicated tuning session. We are aware of the fact that the extreme diversity of OSNs content and the continuing evolution of communication styles create the need of using several datasets as a reference benchmark. We hope that our dataset will pave the way for a quantitative and more precise analysis of OSN short text classification methods.

Table 1: Results of the Proposed Model in Term of Precision (P), Recall (R) and F-Measure (F1) 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%
F1	87%	61%	59%	55%	74%	49%	89%

In order to provide an overall assessment of how effectively the system applies a FR, we look at Table 1. This table allows us to estimate the Precision and Recall of our FRs, since values reported in Table 1 have been computed for FRs with content specification component set to (C; 0:5), where C 2 . Let us suppose that the system applies a given rule on a certain message. As such, Precision reported in Table 1 is the probability that the decision taken on the considered message (that is, blocking it or not) is actually the correct one. In contrast, Recall has to be interpreted as the probability that, given a rule that must be applied over a certain message, the rule is really enforced. Let us now discuss, with some examples, the results presented in Table 1, which reports Precision and Recall values. The second column of Table I represents the Precision and the Recall value computed for FRs with (Neutral; 0:5) content constraint. In contrast, the fifth column stores the Precision and the Recall value computed for FRs with (Vulgar; 0:5) constraint. Results achieved by the content-based specification component, on the first level classification, can be considered good enough and reasonably aligned with those obtained by well-known information filtering techniques. Results obtained for the content-based specification component on the second level are slightly less brilliant than those obtained for the first, but we should interpret this in view of the intrinsic difficulties in assigning to a messages a semantically most specific category.

7. Conclusions

In this work, we have presented an unwanted text message filtering from OSN user walls. The system uses a ML soft classifier to enforce customizable content-based FRs. Furthermore, the flexibility of the system in terms of filtering options is boosted through the management of BLs.

The first task is the extraction and/or selection of contextual features that have been shown to have a high discriminative power. The second task contains the learning phase. As the underlying domain is dynamically changing, the collection of pre-classified data may not be representative in the longer term.

The use of machine learning text classification technique makes classification more automatic and more efficient. The present batch learning strategy, based on the preliminary collection of the entire set of labeled data from experts, permitted an accurate experimental evaluation but needs to be developed to include new operational requirements. Our plan is to solve this problem by investigating the use of on-line learning paradigms able to include label feedbacks from users in future work. The proposed system may have problems similar to those encountered in the specification of OSN privacy settings.

Our plan is to investigate the development of a GUI and a set of related tools to make easier BL and FR specification, for usability of an application.

References

[1] P. W. Foltz and S. T. Dumais, "Personalized information delivery: An analysis of information filtering methods," *Communications of the ACM,* vol. 35, no. 12, pp. 51–60, 1992.
[2] N. J. Belkin and W. B. Croft, "Information filtering and information retrieval: Two sides of the same coin?" *Communications of the ACM*, vol. 35, no.12, pp. 29–38, 1992.
[3] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," 2010P. J. Denning, "Electronic junk," *Communications of the ACM,* vol. 25, no. 3, pp. 163–165, 1982.
[4] J. Golbeck, "Combining provenance with trust in social networks for semantic web content filtering," *in Provenance and Annotation Data, ser. Lecture Notes in Computer Science,* L. Moreau and I. Foster, Eds. 2006
[5] P. E. Baclace, "Competitive agents for information filtering," *Communications of the ACM,* vol. 35, no. 12, p. 50, 1992.
[6] Boykin, P.O., Roychowdhury, V.P.: Leveraging social networks to fight spam. IEEE Computer Magazine 38, 61–67 (2005).
[7] Carminati, B., Ferrari, E.: Access control and privacy in web-based social networks. International Journal of Web Information Systems 4, 395–415 (2008)
[8] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "Rcv1: A new benchmark collection for text categorization research," Journal of Machine Learning Research, 2004.
[9] M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari, "Content-based filtering in on-line social networks," *in Proceedings of ECML/PKDD Workshop on Privacy and Security issues in Data Mining and Machine Learning (PSDML 2010), 2010.* [10] S. Pollock, "A rule-based message filtering system," *ACM Transactions on Office Information Systems,* vol. 6, no. 3, pp. 232–254, 1988.
[10] S. Pollock, "A rule-based message filtering system," *ACM Transactions on Office Information Systems,* vol. 6, no. 3, pp. 232–254, 1988.
[11] Strater, K., Richter, H.: Examining privacy and disclosure in a social networking community. In:

SOUPS '07: *Proceedings of the 3rd symposium on Usable privacy and security.* pp. 157– 158. ACM, New York, NY, USA (2007)

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

[13] J. Park and I. W. Sandberg, "Approximation and radial-basis-function networks," Neural Computation, vol. 5, pp. 305–316, 1993.

[14] V. Bobicev and M. Sokolova, "An effective and robust method for short text classification," in AAAI, D. Fox and C. P. Gomes, Eds. AAAI Press, 2008, pp. 1444–1445.

Author Profile

Ashwini Swami received the B.E. degree in Computer Science and Engineering from SVERI's Institute of Engineering in 2008. During 2009-2012, she worked as the lecturer for polytechnic in same college. And in 2012 she took admission for M.E., and working on project for the same topic.