I.J. Wireless and Microwave Technologies, **2015, 4, 47-61**
Published Online July 2015 in MECS(http://www.mecs-press.net)
DOI: 10.5815/ijwmt.2015.04.05

Available online at http://www.mecs-press.net/ijwmt

Enhanced Techniques for Filtering of Wall Messages over Online Social Networks (OSN) User Profiles

Nikhil Sanyog Choudhary[a], Prof. Himanshu Yadav[b], Prof. Anurag Jain[c, *]

[a] *Department of Computer Science, R.I.T.S, Bhopal, India*
[b] *Department of Computer Science, R.I.T.S, Bhopal, India*
[c] *Department of Computer Science, R.I.T.S, Bhopal, India*

Abstract

Online Social Networks enables various users to connect and share their messages publicly and privately. On one hand it provides advantages to the users to connect and share but on the other hand it provides disadvantage of being attacks or post messages which contains negative or abuse words. Hence OSN provides various filtering rules for security against these wall messages. Although there are various filtering rules and classifiers implemented for the filtering of these users wall messages in popular OSN such as Twitter and Facebook. But in the proposed methodology not only filtering of these wall messages is done but the categorization of normal or negative messages are identified and hence on the basis users can be blacklisted. The proposed methodology is compared with FCM and SVM for clustering and classification of messages. This approach efficiently categorizes the messages but restricts for generating filtering rules and blacklist management. Thus the approach with FCM and J48 first initializes clustering using FCM followed by generation of rules using J48 based decision tree. Hence on the basis of the rules generated message are classified and message which doesn't contain attacks is then filtered on the basis of dictionary which contains a list of abuse words. The methodology is implemented by applying FCM and SVM and a comparison is done with FCM and J48 for the performance on the basis of accuracy to detect abnormal messages.

Index Terms: OSN, SVM, FCM, J48 Classifier, Filtering Rules.

1. Introduction

Internet today can be considered as largest information storage providing ease to the users to solve various queries, gain knowledge, messaging, build networks etc. The user's who continuously access the internet also share, disseminate and communicate numerous type of information and data among them. This information is

* Corresponding author. Tel.: +911127582080
E-mail address: [a] sanyog.nikhil@gmail.com, [b] himanshuyadav86@gmail.com, [c] anurag.akjain@gmail.com

exchanged via emails, messages that may be in any form like audio, text, video, images etc. Various users nowadays access online social networks for building social networks which have basis upon similar interests, likes and dislikes and even for messaging purposes.

1.1. Online Social Networks

OSN's are the platforms running on internet which provides space for user's for sharing of multimedia information among neighboring users. With the help of social networks the present internet generation is able to maintain interaction with the technology as well as other people. OSN's can be considered as a amalgamation of technological, economical and social forces fulfilling the need of the users for building social networks, relations etc. over the internet [1].

Various OSN's like Facebook, Twitter, Instagram, Linked In, Google Plus, Youtube etc. contains dynamic characteristics providing various or say several services to the users apart from messaging and sharing information. The Web is generally organized around the content in a large manner whereas online social networks are planned or organized around the users. In an OSN the users trying to access it firstly join a network then publish their profile as required and any content according to the interest and create links for other users with whom they associate or want to share the content/information.

Social network then provides base over the internet for maintaining the social relationships among users and helps the users to find other users having similar types of interests. It also provides platform for publishing of content and providing knowledge which is provided by other users and also shared and endorsed by other users [2]. In online social networks free space and software tools are provided with the help of content management system for the users to create networks, build relationships and create a public or semi public profile.

The users in social networks write or insert their personal information followed by various other types of information in the form of hobbies, interests, passions etc. by providing this personal data user permits other parties to look about their personal information and grow network according to the common interests, likes, thinking etc.

With advancements in such type of process the user also gets vulnerable towards the users who may be fake or inappropriate as according to the other users. As OSN's provide feature of chat, wall posts, messaging, blogging etc. any user can publish or post message to other user. Although with some privacy agreements this can be prevented but up to some extent as some of the OSN's does not properly fulfil the privacy agreements. Such types of messages are termed as SPAM messages. SPAM's are also present in mails which are filtered accordingly by the mail service provider but are not 100% efficient in filtering the SPAM's.

SPAM's are basically annoying messages or mails sent over the network with a purpose to create nuisance to the receiver or the user. In OSN's SPAM is very common as each user has right to look into other users profile and post a comment on his/her profile. Research in the area of spam detection and filtering has focused upon tasks like non stationary data source, sampling bias in the training data etc. but advancements in technology has made SPAM's to evolve over time with more destructive features and technique that prevents them from filtering [3]. SPAM filtering technique should also be capable to block the SPAM contents and adopt learning procedures to match up the SPAM technique. While adopting filtering procedures for SPAM's multiple difficulties are faced in the form of short text's etc. Short messages contains few words which creates challenge for bag of words based spam filters which filters the SPAM's according to its heavy and non reliable content as very less content is found in the message that can be recognized as non beneficial while comparing and searching it on search engines [4].

Adopting filtering rules, concepts and criteria's SPAM messages can be filtered by analyzing the content of the message. The rules can be defined on the basis like in OSNs the same message that is to be posted may have different or multiple meaning and relevance and depends upon the user who is writing the message. The message creator can be selected for applying the filtering rule through imposing various conditions on the message creator.

The filtering rule can be made to block the message posting according to the message content or according to

the user who creates the message. This simply describes that the rules can directly be imposed upon other users for prevention of SPAM's on OSN's through some privacy policy and agreements. The rules can generally be defined as in accordance with or by user itself comprising of user's age, trust level and thereby over the content contained in the message created by the user.

With the help of filtering rules blocking the user from posting the message can also be made possible either permanently or partially by adopting the concepts of blacklisting. Blacklisting generally facilitates the user to block certain users in their profile on the basis of user's message posting habits and also on content in the post or the message.

The blacklist mechanism and the filtering rules can generally be managed by the system administrator directly. Users can specify or mention their requests or give a feedback related with the other user and thereby demand to process the feedback accordingly. Thus through filtering criteria, blacklist mechanism and the user's feedback the system administrator can adopt the strategy to take action accordingly [5]. The concept lying behind blacklist mechanism enables the user to prevent their walls from unauthorized or say inappropriate posting of messages with illegal content and without the user's consent from a particular message creator while maintaining the relationship with the user.

For deploying such type of technique the data rolling over the OSN's in form of messages, images, audio, video etc. is first needed to be recognized, analyzed and generalized according to the content basis through some text mining approach. The data can be made to form clusters of various types and which can then be used to generate the rules and adopt blacklist concepts accordingly. Clustering comprises of objects which are classified into different groups. In this the data set is partitioned into subsets (clusters). Basically analysis of cluster refers to an extensive assortment of methods in which dataset X partitioned into sub datasets c and these partitioned clusters are disjoint. Partitioned clusters are defined as non-fuzzy or hard c-partition of X [6].

According to some definite distance gauge the information in each subset share common traits. Basically clustering can be defined in two ways:-

Hierarchical algorithms: here by using the results of previously conventional clusters, consecutive clusters could be found. There types are as follows-

- Agglomerative ("bottom-up"): In this approach each component belongs to separate cluster are taken and a successively huge clusters formed by merging them.
- Divisive ("top-down"): The whole dataset is divided into small clusters.

Partitioned clustering: This decides all clusters at once. They include:

- K-means and derivatives: In this algorithm n objects are clustered into k partitions on the basis of attributes where $k < n$. The main goal of this algorithm is to locate centres of natural clusters in the information.
- Fuzzy c-means clustering: This is a technique of clustering in which one part of information can belong to two or more clusters. FC-Means is repeatedly used in pattern recognition.

To identify a structure for clustering at technical level complexity arises. Universally there is no proper algorithm that provides a solution to these problems. That's why it is important to inspect the uniqueness of problem and after that apply proper cluster strategy. Clustering doesn't endow with an appropriate classification of unseen samples generated from the equivalent probability allotment, rather aim of clustering to split fixed unlabeled dataset into distinct and fixed set of ordinary veiled data [7].

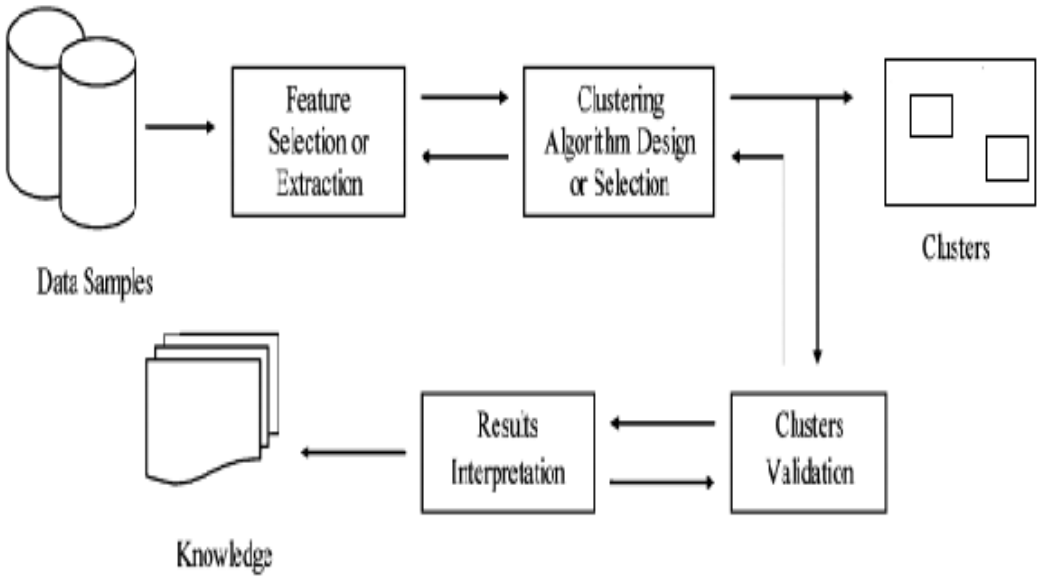

Fig.1. Clustering procedure

1.2. Fuzzy C-Means

FCM being a part of clustering explains the uncertainty of certain item or belonging through a membership function i.e. the data bounds to the clusters through membership function. It is generally a clustering method in which a data may belong to two or more clusters. The FCM works upon the concept of minimization of objective function and with the help of iterative optimization of this objective function fuzzy partitioning is conceded. The fuzzy behaviour of the FCM algorithm is symbolized by the membership function that bounds the data [8].

Taking into account the concepts of machine learning another clustering algorithm is SVM (Support Vector Machine). A SVM based model represents the data items as points in space such that the different data items belonging to different category are divided through a widest gap possible. The new data items are then added into the same divided space and are expected to belong to a category on the basis that which side of the divided gap the data items fall on.

Classification in SVM is generally achieved through formation of hyperplanes. In SVM the multiple categories of data items are divided or grouped with the help of hyperplanes. The basic idea behind hyperplane is that through hyperplane best selection of category separation is achieved by analyzing the largest distance with the nearest or the closest training data item. This results in decrement of the generalization error of the classifier. The separation surface associated with SVM i.e. the hyperplane is generally linear or non-linear [9].

These clustering algorithms enable to analyze the clusters and thereby processed to form decision tree.

The following content in the paper is divided accordingly. Heading 2 Related Work gives the glimpse of Survey of various research papers used for better understanding of topic. Followed by Proposed Methodology explaining algorithms used and the techniques under the heading 3. Heading 4 gives the Result analysis in the form of Precision, Recall and F-Score Values. Following is the Conclusion of the paper in Heading 5. References and Author's profiles are added at last.

2. Related Work

In the research Marco Vanetti [5] et. al. researched and found that the OSN's needs to be modified or say should be adaptive towards the messages that are posted on the user walls or over the private space provided by an OSN. This is required to evade or restrict publishing of unwanted/bad content in the form of posts and messages over the wall. The system proposed by them explains the concepts in which the user is allowed to have direct control over the messages and the content posted on the users walls provided by OSN. This is implemented through filtering criteria's, approach of content based filtering, ML based soft classifier etc. These approaches can be customized according to the requirement and the usage. Strategies behind their research comprised of analyses of content and its selection on the basis of relative features that contains unwanted properties in the message.

Their approach comprised of a learning strategy of labeling according to its content which thereby is used for filtering purposes. The filtering criteria's depends upon various types of factors which are in the form of relationships, trust level, interests. The criteria's have multiple base factors. The user's also faces influence with other social network user's which may be according to interests, likes, dislikes on various agendas and topics etc. In the investigation J. Tang [10] et.al. remarked a problem of topic based social influence analysis. They proposed a distributed learning algorithm under Map reduce programming model. Here author proposed a new model Topical Affinity Propagation (TAP) approach to explain the difficulty using a graphical probabilistic model. The social influence analysis problem creates confronts like controlling of both node-specific topic allocation and network structure which enumerate social influence.

A user's influence on others not only depends on their own topic distribution but also relies on what types of social relationships they have with others. The objective is to intend an incorporated move to utilize local characteristics i.e. topic distribution and the global structure which is in form of network information for social influence examination. For such influences to be examined and predicted F. Liu [11] et. al. suggested and explained the technique behind the collaborative filtering (CF) incorporated into social networks information. The basic underlying idea behind their proposed strategy is to increase the performance and the efficiency of collaborative filtering (CF). The CF strategy generally works on the suggestions of the similar type of neighbors and of the user's for a variety of relation, data exchange, information etc. in a social network. Limitations associated with CF are overcome or resolved by increasing the efficiency by recommending the user to others having similar taste or interests those who may be even in neighbors of strangers.

Implicit and explicit information in multiple type of users from social network is used for accurate prediction through facilitating the users to connect and interact with others in their social network thereby using it for exchange of information, business recommendations etc. this also generates a sensitive area of various privacy related issues while exchanging such type of information. O. Kafali [12] et. al. presented PROTOSS which is capable of predicting and detecting various privacy violations. The privacy statements proposed by OSN's sometimes does not fulfill the privacy concerns of the user's. The tool is based upon the relations along with user's privacy agreements and domain based semantic data.

Tool proposed is capable of analyzing and generating hypothetical future state which may be generally the possible future violations. The model presented can predict and inform the user regarding the possible privacy violations that may occur by its own which may even be unknown to the user or may not have described in privacy agreement.

These violations somehow affect the social profile of a user as even after privacy agreements, trust levels etc. the user can become victim of spam messages. P. Oscar Boykin [13] et.al. gave that Social networks are also useful for reviewing the trustworthiness of outsiders. They proposed an anti-spam tool being automatic in nature and capable of exploiting the properties of social networks. With the help of the tool the unsolicited commercial e-mail (spam) are distinguished with the messages that are related or associated with people within the user's network. This technique is predicated on recognizing the unique characteristics inherent to social networks.

They use the quantitative definition of the clustering coefficient that involves counting the fraction of a node's neighbors that are also each other's neighbors. These kinds of approaches thereby aroused the need of filtering criteria's and rules with the help of concepts of text mining. For such needed instances V. Bhujade [14]

et.al. explained and presented automatic extraction of association rules from textual documents with the help of a text mining technique naming it as EART (Extracting Association Rules from Text). Their proposed technique discovers association rules between the important keywords thereby document labeling. The system basically looks towards the statistical distribution of words in the document leaving the occurrence order of words. It is based on TF-IDF information retrieval scheme which picks the keywords required for generation of association rules.

They also gave GARW algorithm for generation of association rules based on weighting scheme. For such kind of efficient system over the web on OSN's needed the concept for the machines to adopt the features of efficient retrieval of information from the network which is relevant in nature and is required. For the machines capability of extracting the information M. Chau et. al. [15] explained the observable facts about searching for relevant or necessary information from the web and filtering out or leaving the inappropriate information. The same concept can be used for message filtering in OSN's in which the process involved can be made machine dependent and managed by using machines intelligence. The proposal presented here explains content based and link based features used for machine learning algorithms and filtering out irrelevant information.

These types of extraction of relevant information are configured through access control policies. The need for such type of policies is necessary because according to the research by M. Madejski [16] et. al. deployment of Access control polices by system administrator still are not fully efficient. The privacy setting's provide by OSN are unable to match user's need while sharing the content. According to the research followed by their application results revealed that approx one of the users in 65 case studies is affected by a privacy violation due to non proper management of privacy settings. To prevent such privacy violations the access control policies or privacy settings can be further specified giving the user direct control over his/her wall. The user's can be provided with the content type that is about to post on his wall followed by the permission of the user to allow the post or discard it. This can be achieved through analysis of text through text mining techniques. N. Kanya [17] et.al. explained that Text mining basically discovers patterns in unstructured text. The process in the approach looks for specific item-sets in the documents which are applied for information extraction (IE).

With the help of DISCOTEX (Discovery from Text Extraction) framework they extracted interesting relationships by converting the text in the documents into structured data.

For the efficient text mining approach the data can be divided into clusters for easy recognition of irrelevant and unwanted data thereby increasing the future efficiency. But in OSN's this data can be very large according to the number of user's and the content. Therefore considering large datasets A. McCallum [18] et.al. suggested that due to possibility of millions of data points and their existence in various dimensions which represents thousands of clusters efficient clustering technique needs to be adopted. Their research on high dimensional datasets proposed technique which approximately measures the distance dividing the data efficiently into overlapping subsets. These are termed as canopies.

Clustering is then followed through process of measuring distance just in between the points that come or lays in common canopy. This has been applied to solve the concepts related to reference matching while describing the descriptions of the objects. The proposed approach here can be a used with various clustering algorithms. For cluster analysis Rui Xu [7] et.al. explained that through clustering algorithms data exploration and analysis of cluster is made possible for examining unlabeled data through hierarchical structure construction and set of groups. The research followed the steps involving pre processing, solution validity, evaluation, algorithm development etc.

It has been formalized in the result that no clustering algorithm solves all the related problems. Algorithms generally base upon some assumptions and also favor some types of biases. A clustering algorithm should be able to generate arbitrary shapes of clusters, handle large datasets, should be able to handle new data etc.

Various types of clustering algorithms includes like K-Means, SVM, Fuzzy C-Means, Hierarchical Clustering approach etc. Weiling Cai [19] et.al. in their research described Fuzzy c-means (FCM) algorithm. The concept here is deployed upon image segmentation proposing a Fast Generalized Fuzzy c-means clustering algorithms (FGFCM) which is capable of enhancing clustering performance. The proposed algorithm is a combination that includes both enhanced FCM and fast FCM making FGFCM more efficient, robust and fast.

FGFCM integrates local spatial and gray information of images generating fast clustering in which attributes depends upon number of gray levels and not on size of image. Fuzzy C-Means algorithm thereby enhances clustering concepts. Furthermore A. Banumathi [20] et.al. explained that for discovering patterns in large dataset clustering is being used in data mining application. They analyzed Fuzzy C-Means algorithm proposing that the resultant cluster quality depends upon initial seed while selecting it in sequential or random manner. Fuzzy C-Means performs initial operation of clustering with the help of K-Means thereby calculating the degree of membership.

They compared Fuzzy C-Means with K-Means and removed the drawback related with K-Means in Unique Clustering with Affinity Measure Clustering algorithm (UCAM) and evolving it as Fuzzy-UCAM. They gave that in real time large databases initial seeds and number of clusters cannot be predicted accurately which is required by Fuzzy C Means algorithm for its initiation.

Therefore in their proposed approach Fuzzy-UCAM algorithm is used for clustering without providing number of clusters and initial seed. The UCAM and Fuzzy-UCAM given by them generates membership function and is used for clustering of data which results in reduction of overheads related with the cluster sixe and initial seeds. The approach also sets a threshold value for obtaining unique clustering thereby increasing the scalability and decreasing the clustering error ensuring clustering mechanism taking place in time and accuracy of clusters is not lost.

Another clustering algorithm used for classification defined by S.V.N. Vishwanathan [9] et.al. explains fast iterative algorithm to identify Support Vectors on a given set of points. In their proposed methodology candidate Support Vector is used to set a greedy approach which is then adopted for selecting the points required for inclusion in the candidate set. The points contained in the candidate set at times blocks new point's addition. This is resolved using the backtracking approach.

Thereafter convergence procedure is initialized and enhanced through points which are from opposite classes and are nearest to each other. Following the increment or trimming of candidate Support Vector set. Basically their proposed approach surmounts iterative algorithms on taking into account the kernel computations.

Their proposed algorithm does not have numerical instabilities and free from round off errors without using the kernel cache while reusing the kernel computations. While summarizing their methodology the concepts of clustering algorithms enables to filter spam's and messages while being resourceful in their approaches.

In the paper Q. Wang [3] et.al. enlighten that using SVM classification it is possible to separate emails in the categories of valid and fake directly depending upon their classification properties. The properties are generally analyzed through online learning algorithms. Their approach is capable of analyzing the spam messages that have the capability to advance with time and technology. But also has difficulty for labeling of messages because of abundance of information content over the web thereby requiring initial training. The technique generally offers advancements in methods related to spam filtering over random selections and also over delay retraining methods. Thus with the help of active learning the technique of spam filtering attain generalization accuracy while expressing the number of labeled examples. Whereas online learning algorithm makes retraining process fast.

3. Proposed Algorithm

The proposed methodology with FCM and SVM implemented here consists of following steps:

1. Take an input dataset of online social networks such as Facebook, twitter or co-authorship.
2. Apply Clustering on the input dataset to generate two clusters using FCM algorithm.
3. For each of the generated cluster it is classified using SVM.
4. Classification of the data is done as attack and normal.

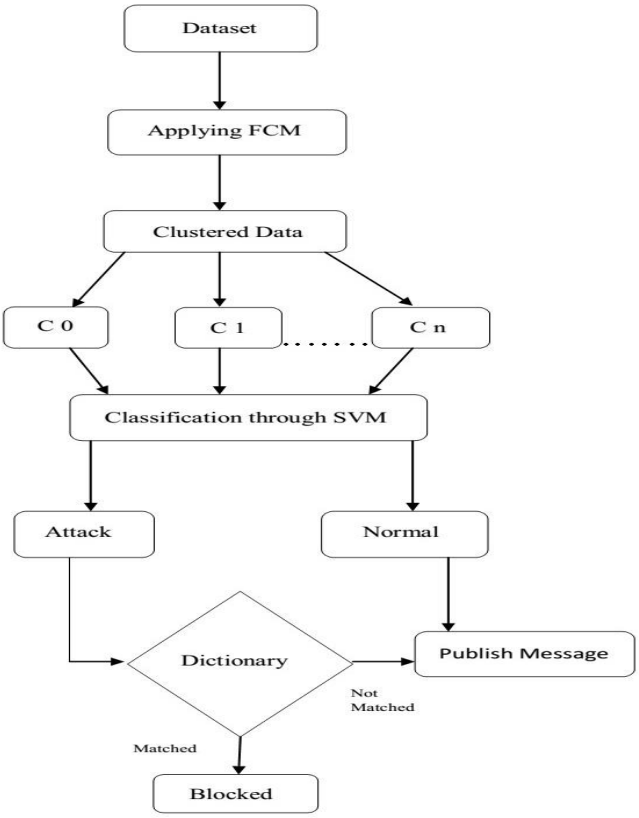

Fig.2. Flow Chart of the Proposed Methodology with FCM and SVM

But here this technique although being efficient in classifying the dataset as normal and attack messages, rules generation and blacklisting of user is not possible. This required a need to fulfil these criteria's.

Therefore proposed methodology comparison with FCM and J48 implemented here consists of the following steps:

1. Take an input dataset of Online social networks such as Facebook, Twitter or co-authorship (Para a).
2. Apply Fuzzy C-means clustering the OSN Dataset for the grouping of similar and dissimilar data (Para b).
3. Now Apply J48 based Classification algorithm on the clustered dataset to generate a decision tree (Para c).
4. Each of the users wall message is compared with the fuzzy rules generated (Para d).
5. The messages are then filtered using the dictionary which contains negative words (Para e).

- Para a

Here the input dataset is a collection of number of wall messages containing negative and positive words. The online social networks allows various users to post their messages on the other's wall but these wall messages may contains some negative words which is not publicly to other users and also the chances that the user is a blacklisted user. Here we have collected messages from various sources such as Co-authorship dataset and Facebook Dataset and Twitter Datasets which can be passed an input the algorithm for filtering.

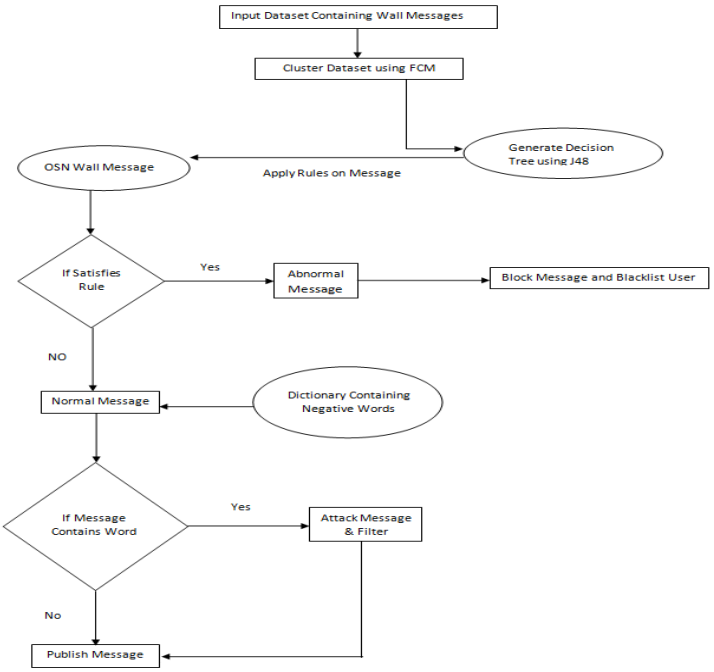

Fig.3. Flow Chart of the Proposed Methodology with FCM and J48

- Para b

Fuzzy C-Means (FCM) is a clustering approach in which one piece of data can belong to two or more different clusters. The approach is regularly used in pattern recognition. Among the fuzzy clustering methods, fuzzy c-means (FCM) algorithm is the most popular method used in image segmentation because it has robust characteristics for ambiguity and can retain much more information. Although the conventional FCM algorithm works well on most noise-less images, it has a serious limitation like: it does not incorporate any information about spatial context that cause it to be responsive to noise and imaging artefacts. To compensate for this shortcoming of FCM, the observable approach is to smooth the image before segmentation. Nevertheless, the conservative smoothing filters can result in loss of important image details, especially image boundaries or edges. More importantly, there is no way to rigorously control the trade-off between the smoothing and clustering. The algorithm is an iterative clustering method that produces an optimal c partition by minimizing the weighted within group sum of squared error objective function JFCM:

$$J_{FCM} = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^{q} d^{2}(x_k, u_i) \quad (1)$$

where $X = \{x_1, x_2, \ldots, x_n\} \subseteq Rp$ is the data set in the p-dimensional vector space, n is the number of data items, c is the number of clusters with $2 \leq c < n$, u_{ik} is the degree of membership of x_k in the i^{th} cluster, q is a weighting exponent on each fuzzy membership, v_i is the prototype of the centre of cluster I, $d^2(x_k, v_i)$ is a distance measure between object x_k and cluster centre v_i. Let Vi be the set of vector values in the data points Pi.

a. Initialize membership value U from the set of data point Pi randomly.
b. After k-step calculate the centroid C=[cij] up to the number of clusters using

$$c_j = \frac{\sum_{i=1}^{n} (u_{ij})^m x_i}{\sum_{i=1}^{n} (u_{ij})^m} \tag{2}$$

Where m is the fuzzy parameter and n is the number of data points.

c. After each iteration fuzzy membership is updated using,

$$u_{ij} = \frac{\left(\frac{1}{\left\| x_i - c_j \right\|} \right)^{\frac{1}{(m-1)}}}{\sum_{j=1}^{n_c} \left(\frac{1}{\left\| x_i - c_j \right\|} \right)^{\frac{1}{(m-1)}}} \tag{3}$$

d. Stop the fuzzy C-means algorithm if the value of member ship is less than the previous membership, |Uk-Uk-1| < epsilon

- Para c

The clustered dataset is then passed to the J48 classification algorithm for the classification of data. J48 is classification algorithm which generates a decision tree on the basis of which rules are generated. J48 is based on C4.5 classification algorithm which generates binary tree.

- J48 Algorithm:

INPUT:
D //Training data
OUTPUT:
T //Decision tree
DTBUILD (*D)
{
T=φ;
T= Create root node and label with splitting attribute;
T= Add arc to root node for each split predicate and label;
For each arc do
D= Database created by applying splitting predicate to D;
If stopping point reached for this path, then
T'= create leaf node and label with appropriate class;
Else
T'= DTBUILD(D);
T= add T' to arc;
}

While building a tree, J48 ignores the missing values i.e. the value for that item can be predicted based on what is known about the attribute values for the other records. The basic idea is to divide the data into range

based on the attribute values for that item that are found in the training sample.

Hence a decision tree is created from the J48 Classification algorithm.

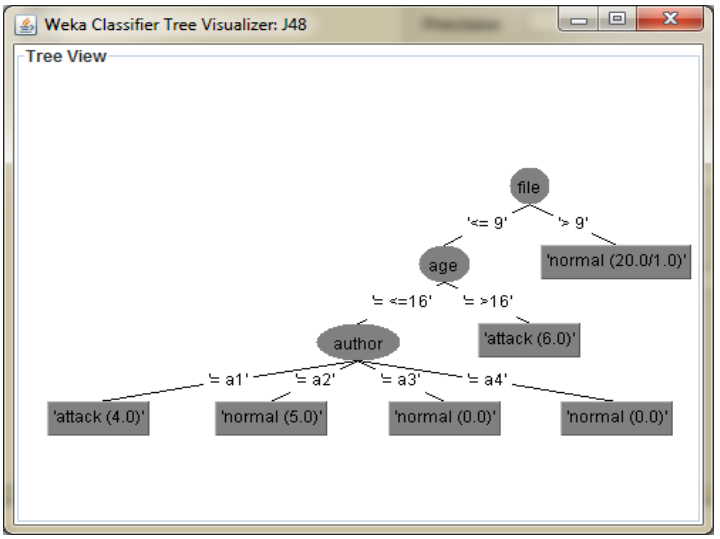

Fig.4. Decision Tree created using J48

- Para d

Each of the users wall message is then compared with the generated fuzzy rules and if the rules are satisfied then the message contains attack and hence both the message is blocked as well as the users is blacklisted.

Table 1. Fuzzy Rules Generation

Fuzzy Rules Generation after Applying J48 using Decision tree
Rule-1

if (file) <= '9' then
if (age) <= '16' then
if (author) == 'a1' then
Message contains Attack
if (author) == 'a2' || 'a3' || 'a4' then
Message is Normal

Rule-2
if (file) >'9' then Message is Normal
Rule-3

if (file) <= '9' then
if (age) > '16' then
Message Contains Attack

- Para e

Here for the message that doesn't contains attacks is then filtered based on the dictionary which contains a number of attacks or negative words.

Here in this technique rules generation and blacklisting of user is possible after clustering and classification of dataset.

But while comparing both the approaches on efficiency terms SVM classification is better than classification using J48 but for further approach classification through J48 is feasible.

4. Result Analysis

Table 2. Values of Precision

Results of Precision		
No. of Messages	**FCM_SVM**	**FCM_J48**
10	0.892	0.82
20	0.91	0.85
30	0.92	0.86
40	0.926	0.868
50	0.94	0.87
60	0.95	0.89
70	0.953	0.893
80	0.96	0.9
90	0.968	0.91
100	0.97	0.93

Table 3. Values of Recall

Results of Recall		
No. of Messages	**FCM_SVM**	**FCM_J48**
10	0.82	0.74
20	0.827	0.76
30	0.84	0.79
40	0.85	0.793
50	0.867	0.81
60	0.88	0.82
70	0.89	0.84
80	0.9	0.847
90	0.92	0.86
100	0.93	0.87

Table 4. Values of F-Score

Results of F-Score		
No. of Messages	**FCM_SVM**	**FCM_J48**
10	0.854	0.78
20	0.866	0.802
30	0.878	0.823
40	0.886	0.828
50	0.902	0.838
60	0.913	0.853
70	0.92	0.865
80	0.929	0.872
90	0.943	0.884
100	0.949	0.899

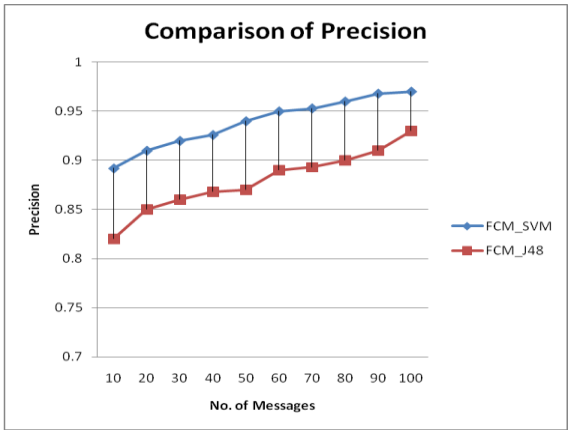

Chart 1. Comparison of Precision

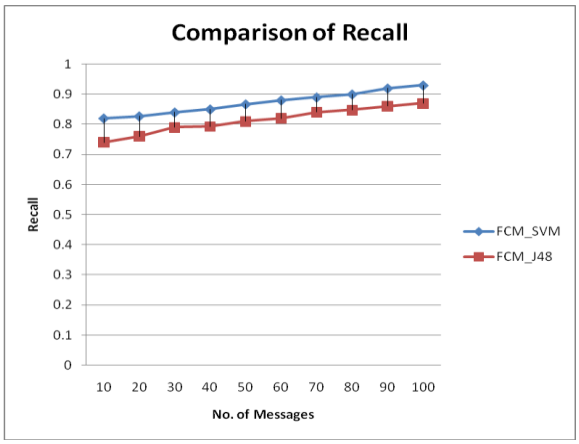

Chart 2. Comparison of Recall

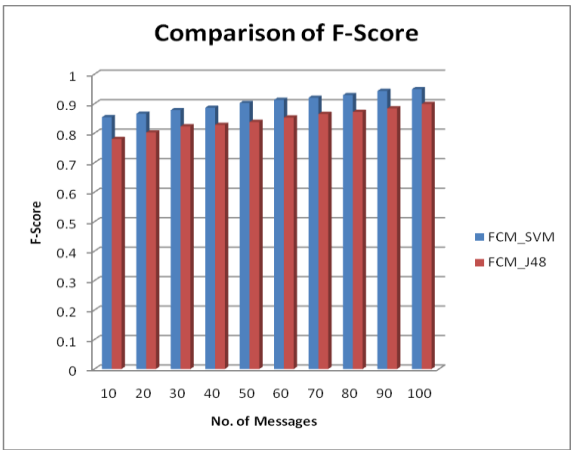

Chart 3. Comparison of F-Score

5. Conclusion

The research has focused on the analysis of various techniques for message filtering in OSN. The messages are clustered using Fuzzy C- Means thereby classifying them on the basis of the content in the messages which needs to be blocked for posting. Classification using SVM efficiently categorizes the messages. Concept of addition of new words is made possible that can be added by the users which are needed to be blocked.

Experiments revealed that the content of the message will be compared with the dictionary of attack words and can be blocked and any of the new messages by the user can be directly categorized for blocking the content.

On comparing the results with J48 classification scheme the messages are efficiently organized in SVM classification scheme. The development of a user and administrator friendly GUI and a set of related tools is done to make easier FR (Filtering Rule) specification which is a direction we planned and adopted. The proposed technique is efficient as compared to the existing technique while filtering the messages.

References

[1] Walter Willinger, Reza Rejaie, Mojtaba Torkjazi, Masoud Valafar and Mauro Maggioni, "Research on Online Social Networks: Time to Face the Real Challenges", 2009.

[2] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel and Bobby Bhattacharjee, "Measurement and Analysis of Online Social Networks", ACM, 2007.

[3] Qiang Wang, Yi Guan and Xiaolong Wang "SVM-Based Spam Filter with Active and Online Learning", 2003.

[4] Gordon V. Cormack, José María Gómez Hidalgo and Enrique Puertas Sanz "Spam Filtering for Short Messages", ACM, 2007.

[5] M. Vanetti, E. Binaghi, E. Ferrari, B. Carminatti and M. Carullo "A System to Filter Unwanted Messages from OSN User Walls", IEEE Transactions on Knowledge and Data Engineering, 2013.

[6] James C. Bezdek, Robert Ehrlich, William Full, "FCM-The Fuzzy C-Means Clustering Algorithm", Computers & Geosciences Vol. 10, No. 2-3, pp. 191-203, 1984.

[7] Rui Xu and Donald Wunsch "Survey of Clustering Algorithms", IEEE Transactions on Neural Networks, 2005.

[8] M.S. Yang "A Survey of Fuzzy Clustering", Mathl. Comput. Modelling, 1993.

[9] S.V.N. Vishwanathan and M. Narasimha Murty "SSVM: A Simple SVM Algorithm", 2002.

[10] Jie Tang, Jimeng Sun, Chi Wang and Zi Yang, "Social Influence Analysis in Large-scale Networks" KDD'09, June 28–July 1, ACM, 2009.

[11] F. Liu and H. J. Lee "Use of Social Network Information to Enhance Collaborative Filtering Performance" Expert Systems with Applications, Science Direct, Elsevier Ltd.-2009.

[12] O. Kafali, A. Gunay and P. Golum "Detecting and Predicting Privacy Violations in OSN" Distributed Parallel Database, Springer Science, 2013.

[13] P. Oscar Boykin and Vwani P. Roy chowdhury. "Leveraging social networks to fight spam" Computer, 38(4):61–68, 2005.

[14] Vaishali Bhujade and N. J. Janwe "Knowledge Discovery in Text Mining Technique Using Association Rules Extraction", International Conference on Computational Intelligence and Communication Systems, 2011.

[15] M. Chau and H. Chen "A Machine Learning Approach to Web Page Filtering using Content and Structure Analysis" Decision Support Systems, Science Direct, Elsevier B.V. 2007, doi:10.1016/j.dss.2007.06.002.

[16] Michelle Madejski, Maritza Johnson and Steven M. Bellovin "A Study of Privacy Settings Errors in an Online Social Network", 2011.

[17] N. Kanya and S. Geetha "Information Extraction -A Text Mining Approach", ICTES 2007.
[18] Andrew McCallum, Kamal Nigam and Lyle H. Ungar "Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching", 2000.
[19] Weiling Cai, Songcan Chen and Daoqiang Zhang "Fast and Robust Fuzzy C-Means Clustering Algorithms Incorporating Local Information for Image Segmentation", 2004.
[20] A. Banumathi and A. Pethalakshmi "Increasing Cluster Uniqueness in Fuzzy C-Means through Affinity Measure", Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, IEEE, 2012.

Authors' Profiles

Nikhil Sanyog Choudhary is currently pursuing M.Tech. in Computer Science from Radharaman Institute of Technology and Science, Bhopal, M.P. India. He has previously submitted research papers in journals and is always interested to innovate and learn new technologies and study various researches.

Prof. Himanshu Yadav is currently working as a professor of Computer Science and Information Technology at Radharaman Institute of Technology and Science, Bhopal, M.P. India. He has submitted multiple research papers in journals and conferences. He has also guided numerous students for M.Tech. with his esteemed knowledge.

Prof. Anurag Jain is currently working as a Head of Department of Computer Science at Radharaman Institute of Technology and Science, Bhopal, M.P. India. He is currently pursuing his PHD. With his vast knowledge has submitted multiple research papers in journal, conferences and transactions. He has been guiding students for a very long period over their M.Tech.

How to cite this paper: Nikhil Sanyog Choudhary, Himanshu Yadav, Anurag Jain,"Enhanced Techniques for Filtering of Wall Messages over Online Social Networks (OSN) User Profiles", IJWMT, vol.5, no.4, pp.47-61, 2015.DOI: 10.5815/ijwmt.2015.04.05