An Improved Approach for Twitter Data Analysis using Clustering and J48 Classification

Ravikant Choudhary
I.T Dept.
Samrat Ashok Technological Institute
Vidisha(M.P), India

Deepak Sain
Asst. Prof. I.T Dept.
Samrat Ashok Technological Institute
Vidisha(M.P), India

ABSTRACT
Social Media Network is one of the main source of data for various event detections. Here in this paper a new and efficient method for the Detection of Traffic in Online Social Network Data is proposed using Clustering and Classification of Data. The Planned Procedure applied here is based on SVM Supervised Learning based Clustering of Similar features of Traffic and then classify the Data using J48 Decision Tree to classify number of events performed in the Twitter Traffic. The Planned Procedure is then compared with the Existing Classification approached such as SVM and Naïve Bayes and C4.5, but the technique is more efficient in comparison.

Keywords
Online Social Media Network, J48, SVM, Naïve Bayes, Real Time Traffic, C4.5, Classification.

1. INTRODUCTION
Social networks have experienced a extraordinary expansion in modern years. Such networks make available a tremendously appropriate space to instantaneously split multimedia information between human beings and their neighbors in the social graph. Social networks make available a prevailing reflection of the arrangement the dynamics of the society and the communication of the Internet creation with both people and technology. Definitely the remarkable growth of social multimedia and client produced content is transforming all phases of the content value chain including manufacture, processing, distribution and expenditure. It also creates and brought to the multimedia sector a novel take too lightly and now dangerous feature of science and knowledge which is social interaction and networking. The significance of this novel quickly developing research area is obviously confirmation by the many related rising technologies and applications like online content sharing services and communities, multimedia statement over the Internet, social multimedia exploration, interactive services and activity, health care and protection applications. It has produced a novel research region called social multimedia computing in which well recognized computing and multimedia networking technologies are brought mutually with promising communal media investigation.

Social networking examinations are altering the approach they converse with others entertain and actually live. Social schmoozing is one of the chief details why more persons have become avid Internet users, people who until the appearance of social systems could not discover importance's in the Web. This is a extremely strong pointer of what is really fashionable connected. Nowadays, users both harvest and devour important numbers of hypermedia content. As social networks will continue to evolve, the discovery of communities, users' interests [1], and the construction of specific social graphs from large scale shared networks will maintain to be a self-motivated investigate challenge [2]. Research in dynamics and trends in social networks may provide valuable tools for information extraction that may be used for epidemic predictions or recommender systems [3-5]. Moreover, their behavior finished online societies is forming a new Internet era where software gratified distribution finished Social Networking Sites (SNSs) is an everyday practice. More than 200 SNSs of worldwide impact are known today and this amount is increasing rapidly. Many of the accessible peak web sites are either SNSs or offer some social networking capabilities. The simplest and most straightforward approach to enable TEM is to apply the traditional TDT solution directly to the Twitter stream. However, this is not desired since the input and output of TDT tasks for traditional event monitoring is in a format which is very different from tweet pre-processing and post-processing the Twitter stream become two necessary steps in order to realize the same standard of detection results as the TDT solutions produce.

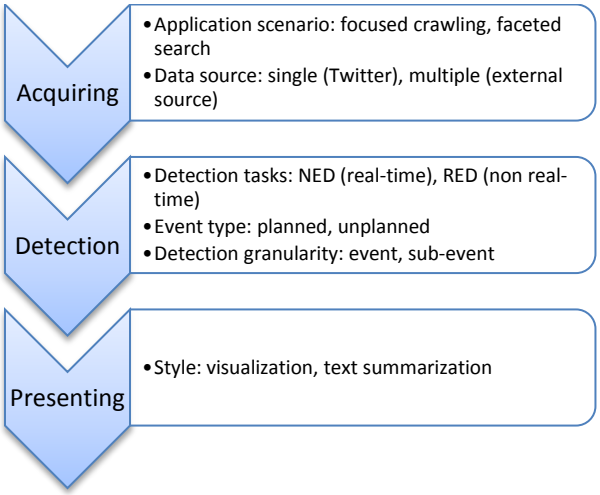

Fig 1: Event Monitoring Pipeline under Twitter

This pipeline is very similar to the production pipeline of traditional. However, rather than producing the news report manually (for example, by news experts), the pipeline of a TEM system emphasizes the automatic identification of newsworthy events from the raw and unstructured user input. This research concludes that the production of event information in a TEM system is achieved by three individual procedures:1) acquiring event information from general public;2) analyzing the raw data for detecting ongoing events, 3) synthesizing the detection results for presenting event summarization to users, as illustrated by Figure 1.

Since traditional media and online media reports are produced by a smaller number of news professionals compared to the normal public, they can be misleading or biased. User

generated event descriptions on online social media, such as Microblogging services can provides more comprehensive information [6]. While some researchers have tried to adapt the existing TDT solutions to an online social media scenario, other researchers have explored new ways to undertake event and sub-event identification. This is because traditional TDT solutions are unable to scale to process the massive amount of streaming. With the rising number of real-world events that are initiated and talk about over social networks, event uncovering and following is appropriate a persuasive investigate concern. on the other hand, the conventional methods to event uncovering and event tracking on huge text streams are not appropriate for the reason that of the following difficulties. First, they are not planned to contract with a huge number of small and deafening messages. Second, social networks contain network arrangements such as friends, followers, respond, and re-tweets. Third, social network messages are related with positions which can be either senders' existing positions or event positions. Fourth, each message is also related with a timestamp. Messages frequently control illuminating and appropriate event information on the other hand; conventional text processing methods imagine documents are non-temporal. in addition, given a exacting time frame and a location the client is concerned in events that happened in the given time structure from the chosen region are more expensive than others. Finding localized events has not been well studied yet.

2. LITERATURE SURVEY

In this paper [7], here author present a real-time monitoring scheme for traffic event discovery from Twitter stream analysis. The scheme obtains tweets from Twitter according to numerous search criteria; procedures tweets by concerning text mining methods; and to conclude achieves the classification of tweets. The plan is to allocate the suitable class label to each tweet as associated to a traffic event or not. The traffic detection scheme was utilized for real-time monitoring of numerous regions of the Italian road network permitting for detection of traffic events approximately in real time often before online traffic news web sites. Here they utilized the support vector machine as a classification model and they realized a correctness value of 95.75% by solving a binary classification difficulty i.e. traffic versus no-traffic tweets. Here they were also proficient to distinguish if traffic is reasoned by an outside event or not by solving a multiclass classification difficulty and obtaining correctness value of 88.89% for the 3-class problem in which they have also measured the traffic due to outside event class.

Li et al. present Twevent in [8]. It is a state-of-the-art system detecting events from the tweet stream. The authors use the notion of tweet segments instead of unigram to detect and describe events. Given Twitter messages, Twevent firstly segments each individual message into a sequence of consecutive phrases by using Microsoft Web N-Gram. Then bursty sections are recognized by modeling the frequency of a segment. User frequency of the tweet segments is used to identify the event-related bursty segments. Then, a clustering algorithm is functional to collection event-related sections as candidate events. Wikipedia is utilized to approximately evaluate important and unusual aspects of a candidate event. The system architecture of Twevent is shown in Figure 5. As a result, the proceedings perceived with Twevent are seriously subjective by Microsoft Web N-Gram and Wikipedia, which could hypothetically misrepresent the observation of events by Twitter consumers and also offer less importance to modern events that are not yet reported on Wikipedia.

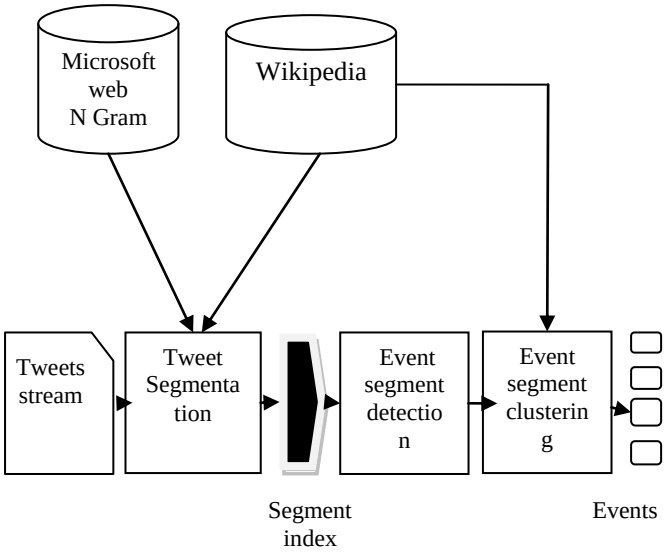

Fig 2:Segment-based event detection system architecture [8]

Chierichetti et al. build a classifier with the volume of tweets and retweet about the broadcasting event for identify important sub-events [9]. This research is based on the fact that users become less social (less volume of retweet) when the event just happen, but quickly back to socializing afterwards when seeing the broadcasting of event. As the supervised machine learning approaches, these solutions require very detailed prior knowledge on the events to be detected. It is compulsory to provide both confident and undesirable instance for training an unbiased classifier on the events.

Kleinberg et al. [10] planned a state mechanism to model the entrance times of pamphlets in a stream. Different states generate time gaps according to exponential density functions with unusual predictable values, and bursty intervals can be discovered from the fundamental condition series. A comparable method by Ihler et al. [11] models a arrangement of amount data using Poisson distributions. However, these methods can only be applied to detect bursty patterns, given a single stream about a certain topic e.g. transportation, sports, politics, etc. News stream provide various daily event-oriented information, which covers almost every area of public life. Topic Detection and Tracking (TDT) is a relatively old research area, which mainly focuses on event identification on news stream. However, each article in a news stream is event-related, which makes it quite different from micro-blogs.

Due to the fast growth of Web 2.0, social media, especially Twitter, essentially different the way people seek information. Nowadays, when hearing or seeing an event, the first reaction of the majority is to post it on Twitter or other social media sites. Such property makes social media a good source for event identification. In addition, relevant tweets about an event can reflect the publics' sentiments and responses to proceedings such as elections and scandals. It is therefore very useful to find popular events and their relevant tweets from Twitter. There have been quite a few works about event identification on twitter in recent years. These studies focus on early event detection, while we study event identification of events from a given segment of Twitter stream in a retrospective manner.

In this paper, they propose an intelligent scheme based on text

mining and machine learning algorithms for real-time detection of traffic events from Twitter stream examination. The scheme subsequent to a practicability learns has been planned and increased from sthe ground as an event-driven communications built on a Facility Concerned with Architecture (SOA) [12]. The scheme uses obtainable technologies based on modern approaches for text analysis and pattern classification. These technologies and methods have been examined, harmony, modified, and integrated with the purpose of build the intelligent scheme. Particularly, they present an experimental learning which has been presented for influential the most efficient among unusual modern methods for text classification. The selected method was included into the absolute scheme and utilized for the on-the-area real-time detection of traffic events.

In this paper, author has using Twitter has numerous benefits over the comparable micro-blogging services. At this point they primary examinations on tweets are up to 140 characters enhancing the real-time and news-oriented environment of the stage. In actual fact the life-time of tweets is frequently very small thus Twitter is the social network platform that is most excellent well-matched to study SUMs associated to real-time events [13]. Second, each tweet can be in a straight line connected with meta-information that represents extra information. Third, Twitter messages are unrestricted, i.e., they are straightforwardly accessible with no privacy restrictions. For all of these causes Twitter is a good foundation of material for real-time happening uncovering and analysis.

3. PROPOSED METHODOLOGY

The planned technique implemented here consists of the subsequent steps:

1. Take a contribution dataset of Online social networks such as facebook, twitter or co-authorship (Para a).
2. Now Smear J48 based Classification procedure on the effort dataset to produce a decision tree (Para b).
3. Produce Fuzzy Rules from the conclusion tree using Fuzzy C-means (Para c).
4. Each of the consumers 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 consumers and also the chances that the consumer 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.

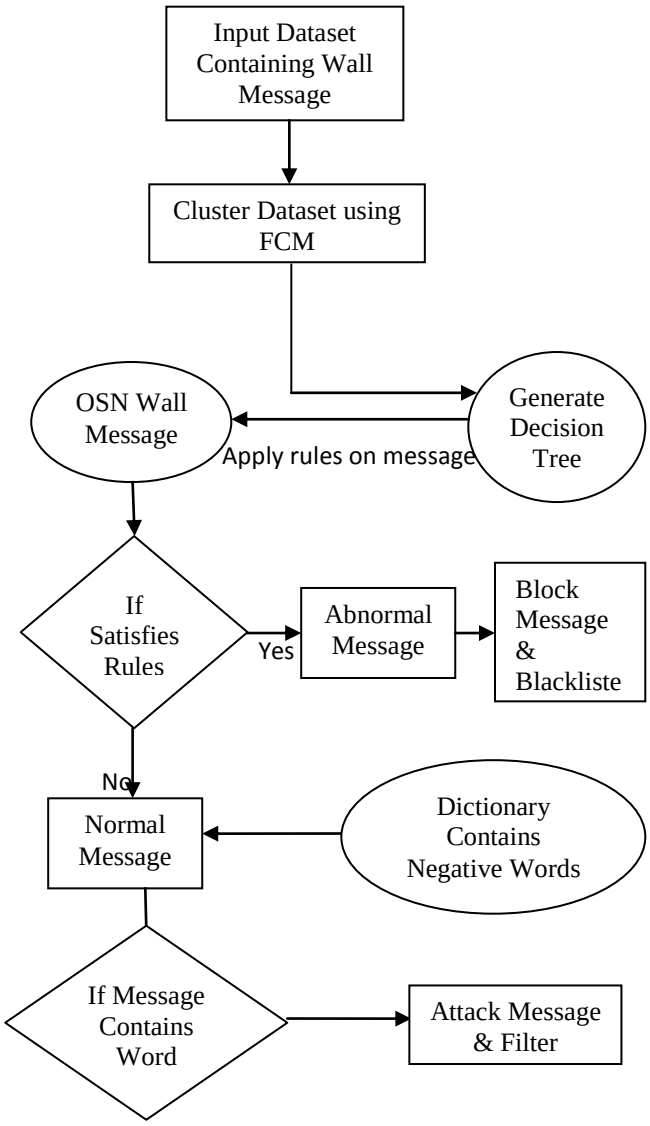

Fig 3: Flow Chart of the planned technique

Para b

The Contribution dataset is then approved to the J48 organization procedure for the cataloging of information. J48 is organization procedure which produces a decision tree on the derivation of which rules are produced. J48 is based on C4.5 classification algorithm which generates binary tree.

INPUT:
D //Training data

OUTPUT:
T //Decision tree

DTBUILD (*D)
{
T=φ;
T= Generate root node and sticker with excruciating characteristic;
T= Enhance arc to root protuberance for each divided establish and
label;
For each arc do
D= Database fashioned by smearing unbearable
predicate to D;
If discontinuing argument grasped for this footpath, then

T'= produce leaf swelling and marker with
Appropriate class;
Else
T'= DTBUILD(D);
T= add T' to arc;
}

While building a tree, J48 ignores the missing values i.e. the assessment for that thing can be forecasted based on what is known about the characteristic values for the other records. The basic idea is to divide the data into variety based on the attribute values for that thing that are found in the instruction illustration.

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

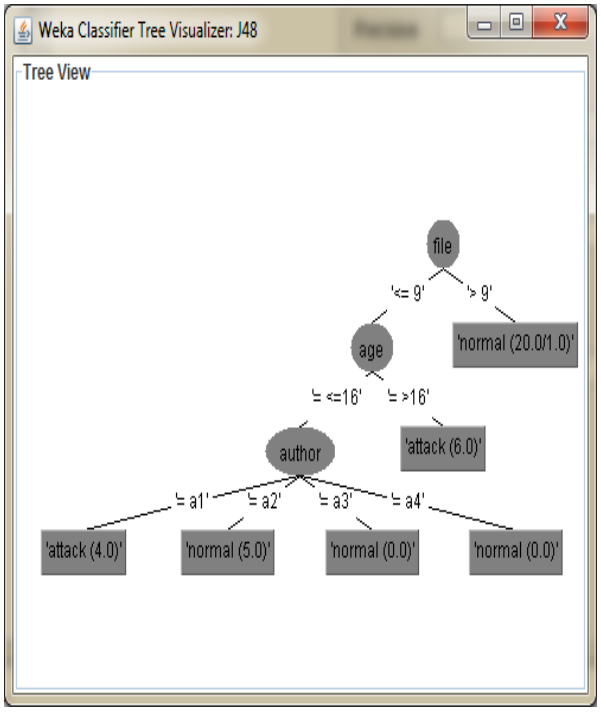

Fig 4: Decision Tree created using J48

Para c
As soon as the decision tree is constructed fuzzy C-means is applied on these classified dataset to generate fuzzy rules.

Fuzzy c-means (FCM) is a technique of clustering which permits one piece of information to go to two or more clusters. It is frequently used in pattern recognition. Along with the fuzzy clustering techniques fuzzy c-means (FCM) algorithm is the most well-liked technique employed in image segmentation because it has strong characteristics for ambiguity and can retain much more information. Although the predictable FCM algorithm efforts well on most noise-less images, it has a serious drawback like: it does not integrate any data about spatial circumstance that reason it to be reactive to noise and imaging artifacts. To reimburse for this shortcoming of Fuzzy c-means the observable approach is to flat the image earlier than segmentation. Nevertheless, the conservative smoothing filters can effect in failure of significant image features particularly image boundaries or edges. More significantly there is no method to thoroughly manage the trade-off connecting the smoothing and clustering. The procedure is an iterative gathering method that concepts an optimum c panel by plummeting the prejudiced within collection sum of formed error impartial function JFCM:

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

where $X = \{x_1, x_2, \ldots, x_n\} \subseteq Rp$ is the information set in the p-dimensional trajectory space, n is the quantity of information items, c is the amount of groups with $2 \le c < n$, u_{ik} is the gradation of membership of x_k in the i^{th} cluster, q is a allowance proponent on each fuzzy association, v_i is the example of the centre of cluster I, d^2 (x_k, v_i) is a aloofness amount amongst object x_k and collection centre v_i.Let Vi be the set of trajectory standards in the information 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

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

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

c. After each iteration fuzzy membership is updated using,

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

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

Fuzzy Rules Generation using FCM

```
**********************************************
                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 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.

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.

4. RESULT ANALYSIS

The Table shown below is the analysis and Judgment of Dissimilar types of Classifiers on the basis of their accuracy. The Proposed Methodology implemented here provides high Accuracy rate in Comparison with other Classifiers.

Table 1. Analysis of Accuracy

Classifier	Accuracy
SVM	95.75
C4.5	95.15
Proposed	97.2

The Table shown below is the analysis and Judgment of Dissimilar types of Classifiers on the basis of their Precision. The Proposed Methodology implemented here provides high Precision rate in Comparison with other Classifiers.

Table 2. Analysis of Precision by Class

	Precision by Class	
Classifier	**Traffic**	Non-Traffic
SVM	95.3	96.3
C4.5	94.4	96.1
Proposed	96.5	97.4

The Table shown below is the analysis and Judgment of Dissimilar types of Classifiers on the basis of their Recall. The Proposed Methodology implemented here provides high Recall rate in Comparison with other Classifiers.

Table 3. Analysis of Recall by Class

	Recall by Class	
Classifier	**Traffic**	**Non-Traffic**
SVM	96.5	95
C4.5	96.1	94.2
Proposed	97.4	96.7

The Table shown below is the analysis and Judgement of Diverse types of Classifiers on the basis of their Final Score. The Proposed Methodology implemented here provides high Final Score rate in Comparison with other Classifiers.

Table 4. Analysis of F1-Score by Class

	F1-Score by Class	
Classifier	**Traffic**	**Non-Traffic**
SVM	95.8	95.7
C4.5	95.2	95.1
Proposed	96.4	96.8

The Figure shown below is the analysis and Judgment of Diverse types of Classifiers on the basis of their accuracy. The Proposed Methodology implemented here provides high Accuracy rate in Comparison with other Classifiers.

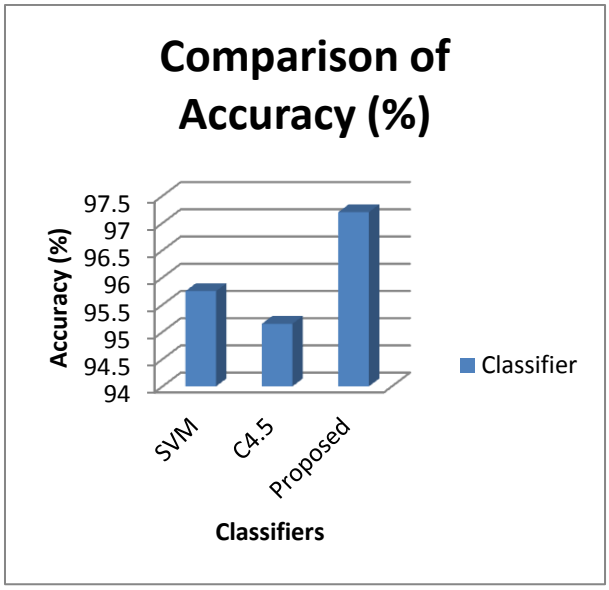

Fig 5:Comparison of Accuracy

The Figure shown below is the analysis and Judgement of Diverse types of Classifiers on the basis of their Precision. The Proposed Methodology implemented here provides high Precision rate in Comparison with other Classifiers.

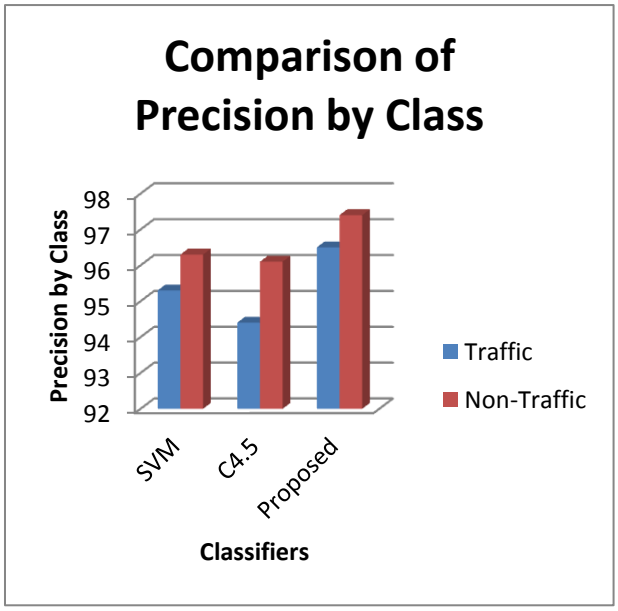

Fig 6:Comparison of Precision by Class

The Figure shown below is the analysis and Judgment of Diverse types of Classifiers on the basis of their Recall. The Proposed Methodology implemented here provides high Recall rate in Comparison with other Classifiers.

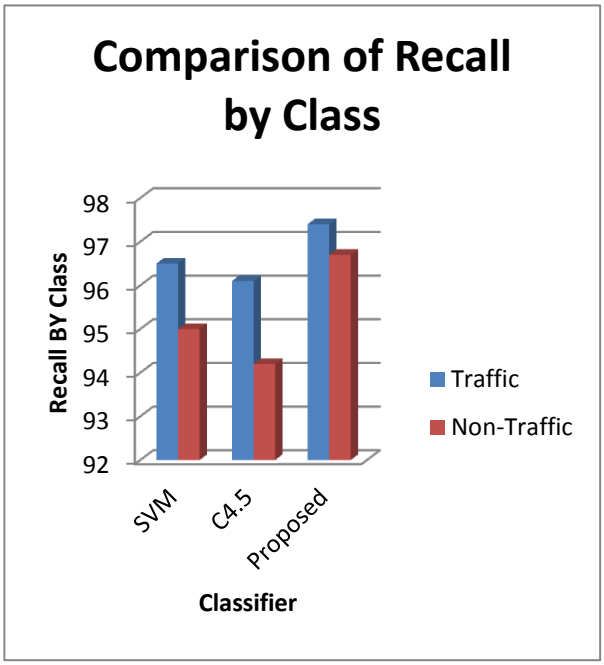

Fig 7:Comparison of Recall by Class

The Figure shown below is the analysis and Judgment of Diverse types of Classifiers on the basis of their Final Score. The Proposed Methodology implemented here provides high Final Score rate in Comparison with other Classifiers.

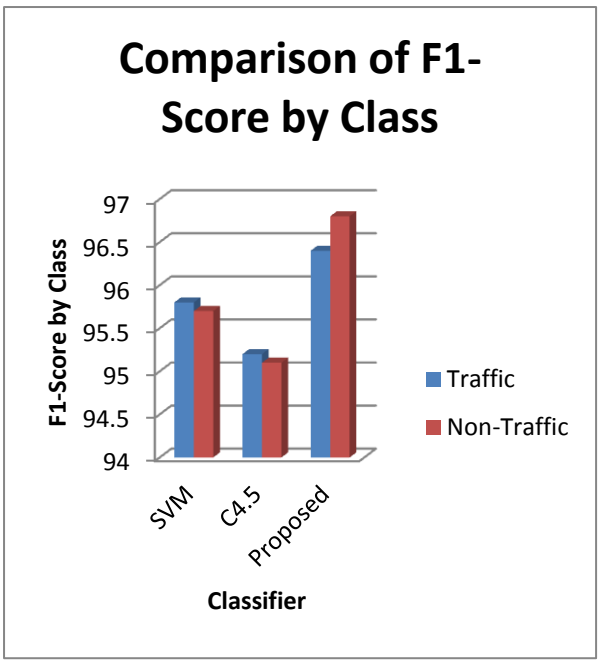

Fig 8:Comparison of F1-Score by Class

5. CONCLUSION

The research has focused on the analysis of various techniques for message filtering in OSN. The messages are clustered using SVM Supervised Learning Approach there by classifying them on the basis of the content in the messages which needs to be blocked for posting. Classification using J48 efficiently categorizes the messages. New words can be added by the user which is 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 expansion of a GUI and a set of connected utensils make easier FR (Filtering Rule) description is a course we planned and adopted. The planned system is efficient as associated to the present technique while filtering the messages.

6. FUTURE WORK

Although the technique implemented here is efficient for the filtering of user wall messages in Online social network, but further enhancements can be done for the improvement of the unwanted words in the dictionary as well as the technique is implemented for the online social tagging on photos as well.

7. REFERENCES

[1] N. Hegde, L. Massoulié and L. Viennot, "Self-Organizing Flows in Social Networks", in Structural Information and Communication Complexity, Lecture Notes in Computer Science, Vol. 8179, pp. 116–128, 2013. (Cited on page 5.)

[2] B.A. Prakash, M. Seshadri, A. Sridharan, S. Machiraju and C. Faloutsos, "EigenSpokes: Surprising Patterns and Scalable Community Chipping in Large Graphs", in IEEE International Conference on Data Mining Workshops (ICDMW), pp. 290-295, Dec. 2009.

[3] T. Sakaki, M. Okazaki, and Y. Matsuo,"Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors", in Proc. of the 19th ACM International Conference on World Wide Web (WWW T10), New York, NY, USA

[4] V. Lampos and N. Cristianini, "Tracking the Flu Pandemic by Monitoring the Social Web", in 2nd International Workshop on Cognitive Information Processing (CIP), pp. 411-416, June 2010.

[5] A. Toninelli, A. Pathak and V. Issarny, "Yarta: A Middleware for ManagingMobile Social Ecosystems", in International Conference on Grid and Pervasive Computing (GPC), pp. 209–220, Oulu, Finland, May 2011.

[6] Y. Demchenko, Zhiming Zhao, P. Grosso, A. Wibisono, and C. de Laat. Addressing big data challenges for scientific data infrastructure. In Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on, pages 614-617, Dec 2012.

[7] Eleonora D'Andrea, Pietro Ducange, Beatrice Lazzerini, and Francesco Marcelloni, "Real-Time Detection of Traffic From Twitter Stream Analysis" IEEE Transactions On Intelligent Transportation Systems, Vol. 16, No. 4, August 2015.

[8] C. Li, A. Sun, and A. Datta. Twevent: segment-based event detection from tweets. In 21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012, pages 155–164, 2012.

[9] Flavio Chierichetti, Jon M. Kleinberg, Ravi Kumar, Mohammad Mahdian, and Sandeep Pandey. "Event

detection via communication pattern analysis". In Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014.

[10] J. Kleinberg. Bursty and hierarchical structure in streams. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 91–101, 2002.

[11] A. Ihler, J. Hutchins, and P. Smyth. Adaptive event detection with time-varying poisson processes. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 207–216, 2006.

[12] The Smarty project SMARTY è un progetto afferente al programma (linea B), ammesso a finanziamento cosi come risulta dalla graduatoria approvata con Decreto Dirigenziale n. 5874 del 10.12.2012.

[13] H. Takemura and K. Tajima, "Tweet classification based on their life-time duration," in Proc. 21st ACM Int. CIKM, Shanghai, China, 2012, pp. 2367–2370.