ISSN:0975-9646

Ranking Web Forum Radical Influential Users Based On Textual Data through Sentiment Classification

J.Jovitha Fastina[1] , Dr.Ilango Krishnamurthi[2]

PG Scholar[1] , Professor and Dean[2]
[1,2]Dept of Computer science and Engineering
Sri Krishna College of Engineering and Technology, Coimbatore, Tamilnadu, India.

Abstract--Web Forums are today's growing society used for several purposes and also being used as a way for practising some harmful acts with concealed ideologies. In any web forum community influential users dominate the normal users through their radical and unwanted thoughts. A threat list is to be maintained by the forum admin, which contains radical or threat words used often by the influential users. The radicalness of the user is obtained through their degree of match of the commented posts with the threat list. Radical users along with their comments are collected separately using collocation theory and Sentiment classification is done to find the exact radical users in the forum. After classification, Page Rank Algorithm is used to Rank the Radical users alone. Through the proposed methodology, the Efficiency in finding the radical users and ranking concept is supposed to provide better results.

Keywords--Sentiment classifier, collocation theory Web Mining, Web forum, Abusers

I. INTRODUCTION

Web is the assortment of billion of documents that is extremely huge, dynamic, diverse and versatile. The web continuous to grow each in large volume of traffic, complexity and size. Web mining is the associate degree rising a part of data processing, that uses the techniques of information mining. However, Web mining doesn't solely mean applying data processing techniques to the information hold on within the Web. There are three main categories of web mining which emphasizes the behaviour of web mining process in world wide web.

Web Content Mining is the method of extracting helpful information from the contents of net documents. Content information corresponds to the collection of facts an web page which is designed to convey them to the users. It carries text, images, audio, video, or structured records like lists and tables. Application of text mining to online page has been the foremost wide research.

The Web also has tremendous success in building communities for users. Identifying such communities are useful for many purposes but security in those communities are not guaranteed. Identified Web communities are 'authoritative' linked together by 'hubs'. Group of individuals with similar interests are identified, who are in the cyber-world would form a "community". Two people are termed as "friends" if the similarity between their Web pages are high. The similarities between them are measured using the features such as text, out-links, in-links and mailing lists with each user in the community. Searching the Web involves two main steps: Extracting the relevant pages based on the query and ranking them according to their quality and usage. Ranking is important as it helps the user look after "quality" pages that are relevant to the queries given by the user. Different metrics have been proposed to rank Web pages according to their quality. Page Rank is a metric for ranking hypertext documents on web based on their quality. The key idea is of ranking is that a page has a high rank if it is pointed to or pointed by many highly ranked pages. Thus the rank of a page depends on the ranks of the pages pointing to it. This process is done iteratively until the rank of all the pages is determined.

II. RELATED WORKS

The Radical influential users in web forums often take over the minds of normal users through their radical thoughts. The solution for the above problem is to Identify the radical users and thus to reduce their interaction in such forums. The following are categories to find the solution for the problem.

A. Content based Radicalness Identification

Content based analysis is used to investigate the content posted by extremist groups on the communities. Affect analysis is the technique used to understand the sentiment or emotion in text towards topics. The paper [1], present a novel agglomerative clustering method to identify cliques in dark Web forums. It Considers each post as an individual entity which accompanies all information about its author, time-stamp, thread etc., having defined a similarity function to identify similarity between each pair of posts as a blend of their contextual and temporal coherence. The similarity function is employed in the proposed algorithm to group similar threads into different clusters that are finally identified as individual cliques.

The clustering of Web opinions today are very challenging due to a certain properties of Web opinions[2] that do not exist in normal documents. Those properties include (1) the messages are less focused, (2) the messages are usually short with certain range of length from few sentences to a couple of paragraphs, (3) different users may use some peculiar terms to discuss the same topic, thus the messages are sparse, (5) there are some noises, many Web opinions do not fall into any categories,. Conventional

document clustering techniques which works well in clustering the regular documents normally do not work well in Web opinions clustering.

In paper[2], the scalable distance-based algorithm for cluster web opinions is proposed. Density-Based Spatial Clustering is an algorithm that filters noise and discover clusters without pre-specifying the number of clusters. Ensemble classifiers use multiple classifiers with each one using different techniques, training sets or feature subsets. Particularly, the feature subset classifier approach in the paper [3] has been shown to be effective for analysis of text patterns. Stamatatos and Widmer used an SVM ensemble classifiers for music performer recognition. They further used multiple SVMs each one trained using different feature subsets. Similarly, Cherkauer used a Neural Network ensemble for imagery analysis in the same field.

B. Network based Radicalness Identification

Network based analysis also known as Link Analysis is used to understand the structure of links between hate promoting websites in the internet. The existence of any link between any two nodes can be due to a hyperlink, friend relationship between them or due to some kind of interaction like a reply to e-mail or a message. Link analysis[6] helps in examining the topology of the network and to characterize the properties of the network.

Social Network Analysis[7] helps in understanding the relationships in a given community and thus analyzing its graph representation. Users are taken as nodes and relations among the users are taken as arcs. Similarly, several techniques have been proposed[4] to extract key members of the network, classify users according to their his relationships within the community, discovering and describing resulting sub-communities in them and also amongst other applications. However, all these approaches leave aside the meaning of relationships among those users. Therefore, analysis based only on reply of mails or posts or messages to measure relationships' strongness or weakness it is not a good indicator. The HITS algorithm[4] has been extensively used in the social network analysis community for different purposes, where the authoritative scores and hub scores can be interpreted as rich information on how different nodes or community members behave. Through this method, one can be able to focus on a specific group of topics to create a some topic-based network, which provides exact information through its lesser density. After that, by using HITS algorithm, key members of the community or network can be extracted from both the network configurations, and results are also radically different[8]. The system[5] consists of an user-friendliness component that uses a human assisted registration approach to get access to Dark Web forums. It also makes use of multiple dynamic proxies and web forum specific crawling parameter settings to maintain forum access. URL Ordering component uses language independent URL ordering features to allow crawling of Dark Web forums across languages. It mainly focuses on group of communities from three different regions such as U.S. Domestic, Middle East, and Latin America/Spain.

C. Limitations

Several network based[4],[5] and content based techniques[1],[2],[3] are used to analyze radical contents in web forums. Network based techniques throw away light on the community structure, key members and topological characteristics of the network. The techniques that are used in the review don't use formal community detection algorithms to detect such communities. Content based techniques analyze the structure and type of content in the forum to gauge an understanding of the purpose of the radical websites like propaganda, fundraising, sharing philosophy etc. Currently, the techniques in the literature perform this classification manually by the help of radicalism experts. This may not be a feasible solution for a security analyst.

This related works, reviewed the state-of-the-art in the area of automated solutions for detecting online radicalization on Internet and social media websites[7]. A novel multi-level taxonomy or framework have been presented to organize the existing literature[7] and present the perspective on the area. The existing studies are categorized based on a multi-level taxonomy on various dimensions such as the social media websites.

III. PROPOSED METHODOLOGY

Due to tremendous growth of user generated contents on social media sites, there are several web communities or web forums. In each forums, there are several comments posted by the users, which dominates the minds of innocent or normal users. There always exists some users who develop some relationship with other members in the forum through their activeness in posting comments, and their comments always receive significant attention of a large community. They mainly do this to capture the innocent normal users through their radical thoughts and to impose their ideologies into them. They play a leading and dominating role in the community, and their comments greatly affect the sentiments of other users. It is very danger to allow those radical users to survive in the forum. Hence they have to be identified and blocked. The words used by radical users are collected as threat list and thus identified in their comments. For radicalness identification among the users, the existing system used collocation theory to find the association among the users. But the system failed to recognize the exact radical users because those users who use the threat words as a phrase or story narration have also been recognized as radicals. Due to this the innocent normal users get affected.

Sentiment Analysis is used on the user posted comments to find the sentiment or opinion of the users on the comment. This Analysis classifies the users as real Radical and normal users. Then the Real Radicals are Ranked according to their number of usage of threat words on the comments. The Proposed system overcomes the drawback of the Existing system thereby using sentiment classifier and thus supposed to improve the efficiency in finding the radical influential users. Initially a web Forum has to be designed and users are allowed to post their comments. From those textual data comments radical users are identified.The steps involved in the proposed

methodology are:(1)The process starts with forum crawling and preprocessing. (2)Then the pre-processed comments are passed through the threat list to find the radical comments along with the user id, timestamp and author. (3)Those users who used the threat words are retrieved along with the users who are associated with them. (4) Sentiment analysis is applied on those collocated users to classify the innocent and radical users. (5) Finally the radical users are ranked using page rank Algorithm.

TABLE
SAMPLE THREAT LIST

	Fight	Freedom
Bomb blast	Kill	Drug
Attack	Jihad	Corruption
Assassinate	Politics	Muslim
Christianity	Country	God

A. Data pre-processing

The process starts with a data pre-processing step in which the URL of the forum home page is passed to the forum, which moves all relevant web pages and eliminates the duplicates. In data pre-processing module, the comments from the web forum are passed through the threat list as given in Table I. The threat list consists of threat words such as kill, bomb blast, attack etc., which are used frequently by radical users. The comments which contains those threat words are extracted as a result of textual Extraction along with the author, user id, timestamp, date. This module forms the basis for second module feature Extraction.

B. Feature Extraction

There exist a friendly relationship between the users interacting in the same thread, and in the perspective of Web forums the term collocation is defined as the association between users co-interacting in same threads. Thus the collocation theory is applied to capture the associativity of different users through their interactions. Consider each pair of users, where *U* is the set of users, and u_i and u_j represent two individual users. In this table, *a* denotes the number of instances or threads in which u_i and u_j have co-occurred, *b* denotes the number of instances in which u_i has co-occurred with all other users in a thread, $(b - a)$ denotes the number of instances in which u_i has co-occurred with all other users except u_j in a thread. Correspondingly, all other values in this table denote the number of instance in which interactions have taken place between the corresponding users. The comments which are identified as radical are collocated as a group along with the user, user id, timestamp and date

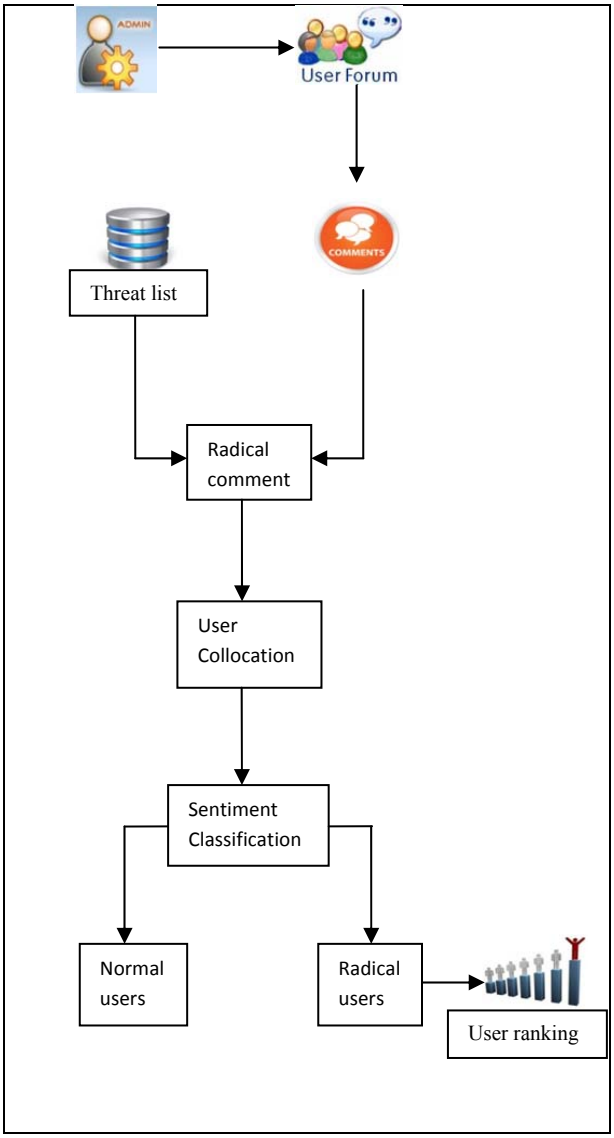

Fig. 1 work flow of the proposed ranking method

. Those users are termed as radically influential users. The users who have posted or replied to the comments against those radical users are also collocated together using the matching of their commented posts through Collocation theory. Thus in this module, only those comments which matches the threat list along with the users are extracted as radicals.

- **Radicalness measure**

Let Ω denotes the set of words in the threat list. A radicalness measure ρ is assigned to each user u_i of the forum being studied, based on the existence of each word Ω_j in each message post $p^i{}_k$ of u_i, where exists(Ω_j , $p^i{}_k$) is a binary function which returns 1 if Ω_j exists in $p^i{}_k$, otherwise 0.

$$\rho(u_i) = \frac{\sum p_k \ \text{posts}(u_j) \sum_j \text{exists}(\Omega_j, p_k)}{\max\{\sum p_k \ \text{posts}(u_i) \sum_j \text{exists}(\Omega_j, p_k)\}}$$

```
Algorithm: LEXICON BASED TECHNIQUE:
Input: SentiWord_Dictionary
Output: Sentimental Analysis (positive, negative or neutral)
BEGIN
 For each threat Ti
{
     SentiScore = 0;
     For each word Wj in Ti that exists in Sentiword_Dictionary
     {
       If polarity[Wj] = blind negation
       {
          Return negative;
       }
  Else
       {
        If polarity[Wj] = positive && strength[Wj] = Strongsubj
        {
         SentiScore = SentiScore + 1;
        }
        Else If polarity[Wj] = positive && strength[Wj] = Weaksubj;
        {
           SentiScore = SentiScore + 0.5;
         }
        Else If polarity[Wj] = negative && strength[Wj] = Strongsubj
        {
        SentiScore = SentiScore – 1;
        }
        Else
        If polarity[Wj] = negative && strength[Wj] = Weaksubj
        {
        SentiScore = SentiScore – 0.5;
        }
          }
          If polarity[Wj] = negation
          {
            Sentiscore = Sentiscore * -1
          }
        }
        If Sentiscore of Ti >0
        {
        Sentiment = positive
        }
        Else If Sentiscore of Ti<0
        {
        Sentiment = negative
        }
        Else
        {
        Sentiment = neutral
        }
        Return Sentiment
      }
      END
```

Figure 2 Lexicon based Dictionary pseudo code for Sentiment Classification

C. Evaluation and Ranking

The threat words may be used normally by the innocent users too. For Example the word 'kill' can be used to represent a person and also an animal as a phrase. So the radical users has to be identified without affecting the innocent users. The Extracted users are evaluated by applying Sentiment classification upon it. Sentiment classification helps to identify both the radical and normal users of the web forums. The opinion or the sentiment behind the comments of the users are obtained through the classification. The result of sentiment analysis provides the true radical users in the web forums.

- **Sentiment Classification**

The Semantic Orientation CALculator (SO-CAL) uses dictionary of collection of words annotated with their semantic orientation(polarity and strength), and incorporates escalation and negation. SO-CAL is applied to the polarity classification task, for the process of assigning a positive or negative score to a text that captures the text's opinion or sentiment towards its specific domain. SO-CAL's performance is dependable across domains and on completely hidden data. The process of dictionary creation, and use of Mechanical Turk to check dictionaries for consistency and reliability must also be considered.

Finally the classified radical users are ranked according to the number of radical comments they have posted. Ranking is done by customized Page rank Algorithm.

- **Page Rank Algorithm**

Threaded discussions among users in the Web forum are used to construct a directed graph by considering each user in the forum as a node, and each user interaction as a directed link. Uni-directional links from all users to the initiator and bi-directional links between each pair of users are established for each thread in the graph. Each user node is initialized with a small value as its page-rank score, and just through Page Rank algorithm, the directed links among them are used iteratively to keep on updating or increasing their rank scores, until a union is achieved.

IV. PERFORMANCE EVALUATION

From Figure 3,The Ranking which has been done in the Existing system(Left), counts the users who all uses the threat words. And hence the threat users count increases though they have been used for normal purpose. This way of ranking includes normal users too. Hence the proposed Ranking(Right) is done after sentiment classification and thus the ranking is based on three factors of sentiment analysis.(i.e)Positive, Negative and Neutral. Thus the performance of the Proposed work exceeds the Existing Ranking by only ranking the Radical users.

Sno	User	Count
1	vidhya saravanavel	7
2	Anwar Sha	5
3	jovitha fastina	3
4	sharanya pravin	2
5	jovitha J	2
6	Yasin Minaza	2
7	Mohammed Rafique	2
8	senthil s	1
9	Deepthi SS	1
10	abirami Boopathy	1
12		

Sno	User	Rank
1	vidhya saravanavel	3
2	Anwar Sha	3
3	jovitha J	1
4	jovitha fastina	1
5	sharanya pravin	0.5
6	kiruba B	0.5
7	abirami Boopathy	0.5
8	Deepthi SS	0.5
9	Yasin Minaza	0.5
10	Mohammed Rafique	0.5
12		

Figure 3 Comparison of Ranking in Existing method(Left) and Proposed Ranking after Sentiment Classification(Right)

V. CONCLUSION AND FUTURE WORK

In the proposed work, radical influential users are identified and ranked. Radicalness measure, collocation theory, sentiment classification and an algorithm based on Page Rank to rank the radically influential users are the major contributions in this project. This work mainly focuses on textual data on the web forums. If any threat is identified in the comment given by the user, the admin has the right to block the user and the user cannot login any more without the authority. This project gives the awareness of abusers who use illegal threat comments in web forums or group threat forming.

REFERENCES

[1] T. Anwar and M. Abulaish "Identifying cliques in dark web forums – an agglomerative clustering approach", in *Proc. IEEE ISI*, Jun. 2012, pp. 171–173.

[2] Christopher C. Yang1 and Tobun D. Ng, "Web opinions analysis with scalable distance-based clustering", in Proceedings of the 2009 IEEE international conference on Intelligence and security informatics, ISI'09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 65–70.

[3] A. Abbasi, H. Chen, S. Thoms, and T. Fu, "Affect analysis of web forums and blogs using correlation ensembles," IEEE Trans. on Knowl. and Data Eng., vol. 20, no. 9, Sep 2008, pp. 1168–1180.

[4] G. L'Huillier, S. A. Ríos, H. Alvarez, and F. Aguilera "Topic-based social network analysis for virtual communities of interests in the dark web", in *Proc. ACM SIGKDD Workshop ISI-KDD*, 2010, Art. ID 9, 2010,pp. 89-102

[5] Tianjun Fu, Ahmed Abbasi, and Hsinchun Chen, "A focused crawler for dark web forums", in Artificial Intelligence Lab, Department of Management Information Systems,2001,pp. 97-104

[6] Chakrabarti, S., Dom, B., Kumar, S., Raghavan, P., Rajagopalan, S.,Tomkins, A., Gibson, D., and Kleinberg, J. Mining the web's link structure. Computer 32, 8 aug,1999,pp. 60 -67.

[7] Ming Yang, Melody Kiang, Yungchang Ku, Chaochang Chiu, Yijun Li "Social media analytics for radical opinion mining in hate group web forums",Journal of Homeland Security and Emergency Management, Vol.no 8, 2011.pp. 1547-7355

[8] H. Chen, W. Chung, J. Qin, E. Reid, M. Sageman, and G. Weimann J. Am.Soc "Uncovering the dark web: a case study of jihad on the web", Inf. Sci. Technol., vol. 59, no. 8, Jun 2008, pp. 1347–1359.