

Abstract— Internet provides its users with anonymity, low publication barrier, low cost of publishing and managing. Online radicalization called cyber extremism and cyber hate propaganda is a growing concern to the society and also of great pertinence to governments & law enforcement agencies The problem of finding out of such hatred videos is proposed. For that there are several tasks: search strategy or algorithm, node similarity computation metric, learning from exemplary poles serving as training data, stopping criterion, node classifier and queue manager. There will implementation of classification algorithm named shark search. There will be comparison of number of words in the language model based comparator, similarity threshold for the classifier and present the results of comparison using standard Information Retrieval metrics such as precision, recall and F-measure. The influential video metadata on YouTube will be studied.

Keywords:- YouTube metadata, Social Network Analysis, online radicalization, Focused Crawler- Best First Search, Frontier Extraction

I INTRODUCTION

Online Radicalization is called Cyber-Terrorism or Extremism is widespread and has become a major and growing concern to the society, governments and law enforcement agencies around the world. Related previous reviews shows that various platforms on the Internet low barrier to publish content, allows anonymity, provides exposure to millions of users and a potential of a very quick and widespread diffusion of message such as YouTube is a popular video sharing website, Twitter is an online micro-blogging service, Face book is a popular social networking website, online discussion forums and blogosphere are being misused for malicious intent. Such platforms are being used to form hate groups, racist communities, spread extremist agenda, incite anger or violence, promote radicalization, recruit members and create virtual organizations and communities. Automatic detection of online radicalization is a technically challenging problem because of the vast amount of the data, unstructured and noisy user-generated content, dynamically changing content. There are several solutions proposed in the literature aiming to combat and counter cyber-hate and cyber-extremism. In this paper we conclude solutions to detect and analyze online radicalization.

II LITERATURE SURVEY

The objectives of the work presented the previous work on to investigate of different techniques used to detect and analyze radical content on the Internet, to perform an in-depth study of these techniques to analyze common trends The high reach ability of videos among users videos are easily accessible to viewers for free, without the need of an account, low publication barriers users need only a valid YouTube account and anonymity their identity is unknown has led users to illegal use YouTube in many ways by uploading malignant content that are illegal [1]. Author (M. Goodwin) works on analyses several hate and extremist groups coming into existence across various countries. He presents an in-depth analysis of their activities, supporters and reasons behind the emergence of these groups. Author (A. Sureka) works on propose an approach based upon the data mining and social network analysis in order to discover hate promoting videos, users and their hidden communities on YouTube.

III SYSTEM ARCHITECTURE OF PROPOSED SYSTEM

A. System Architecture

Below figure1 shows overall working of Focused Crawler for Mining Hate and Extremism Promoting Users, Videos and Communities on YouTube. I have use focused crawler for two different graph traversing algorithms First, Best First Search Algorithm and Second, Shark Search Algorithm. I have use the methodology and solution implementation details for the design and architecture.

Figure1 shows general solution to our system. This fig. include four stages. Collection of Data, Training Data set, Web portal and Testing of Data. In first phase the videos from multiple channels of YouTube like music, sports, Gaming etc will be collected. In second phase the manual analysis on collected data will be performed and make training set. There will collection of around 400 videos promoting hate and extremism that will be used as training m and classified them accordingly. In third phase the web portal will be created. It takes one YouTube channel as a seed positive class channel and extract it’s metadata user activity feeds and profile information using YouTube API.
B. Algorithm Used:
1. Focused Crawler- Best First Search:
The proposed method is a multi-step process primarily consists of three phases, Training Profile Collection, Statistical Model Building and Focused Crawler. We perform a manual analysis and a visual inspection on activity feeds and contextual metadata of various YouTube channels. We collect 35 positive class channels (promoting hate and extremism) used as training profiles. We build our training dataset by extracting the discriminatory features (user activity feeds- titles of videos uploaded, shared, favorite & commented by the user and profile information) of these 35 Channels using YouTube API1. We build a statistical model from these training profiles by applying character n-gram based language modeling approach. We build a focused crawler (best first search) which is a recursive process. It takes one YouTube channel as a seed (a positive class channel) and extract it's contextual metadata (user activity feeds and profile information) using YouTube API.

Algorithm 1: Focused Crawler- Best First Search

Data: Seed User U, Width of Graph w, Size of Graph s, Threshold th, N-gram Ng, Positive Class Channels Up
Result: A connected directed cyclic graph, Nodes= User u

1 for all u 2 Up do
2 D: add(ExtractFeatures(u))
   End
Algorithm BFS(U)
3 while graphsize < s do
4 userfeeds Uf ExtractFeatures(U)
5 score score LanguageModeling(D, Uf , Ng)
6 if (score < th) then
7 U: class Irrelevant
Else
8 U: class Relevant
9 Hashmap Usorted: InsertionSort(U; score)
End
for i   1 to w do
10 Hashmap Ugraph: add(Usorted(i))
End
for all Ug 2 Ugraph do
12 fr = Extract Frontiers(Ug)
13 Hashmap Ucrawler: add(fr)
End
14 for all Ufr 2 Ucrawler do
15 BFS(Ufr)
End

Algorithm 2: Frontier Extraction

Data: User u
Result: Frontiers of a channel
Algorithm Extract_Frontiers(U)

1 usubs u.getSubscribers()
2 ufc u.getFeaturedChannels()
3 ucon u.getFriends()

I find the extent of textual similarity between these metadata and training data by using statistical model build in phase 2 and Ling Pipe API 2. We implement a binary classifier to classify a user channel as relevant or irrelevant. A user channel is said to be relevant (hate and extremism promoting channel) if the computation score is above a predefined threshold. If a channel is relevant, then we further extend its frontiers (links to other YouTube channels) i.e. the subscribers of the channel, featured channels suggested by the user and it's contacts available publicly. We extract these frontiers by parsing users' YouTube homepage using jsoup HTML parser library3. We execute focused crawler phase for each frontier recursively which results a connected graph, where nodes
represent the user channels and edges represent the links between two users.

BFS method has non-binary priority values assigned to each node. The priority values are the similarity score, which is computed by comparing the users’ contextual metadata user activity feeds and profile information with training profiles. Steps 9 and 10 make a list of top w (maximum number of children a node can have) users among relevant users based upon their similarity score, sorted in a decreasing order. Step 11-13 extracts frontiers of a user channel using Algorithm 2. Steps 14 and 15 repeat steps 3-13 for each frontier extracted. We execute this function till we get a graph with desired number of nodes or there is no more node is left to extend.

2. Focused Crawler- Shark Search:
I use a focused crawler for Shark Search Algorithm, an adaptive version of the same algorithm Introduced in M. Hersovici. Shark Search algorithm is different from Best First Search algorithm in a way that it explores frontiers of both relevant and irrelevant nodes. In SSA if the parent of a node is an irrelevant node then the inherited score of the child node is score child *d, where d is a decay factor, an extra input for SSA which directly impacts on the priority of user. This inherited score is dynamic because a node can have more than one parent.

IV RESULT ANALYSIS

As much as the popularity of YouTube for sharing videos is considered a web portal is presented which is used to classify YouTube metadata by using shark search algorithm. This web portal is used to identify hate and extremism promoting videos, users and communities. A series of experiments will be done by varying algorithmic parameters. Social network analysis will be helpful to find hidden communities on YouTube. Comments and videos will be compared with trained data which is done by manual analysis.

REFERENCES
[1] Swat Agawam, Ashish Sureka” A Focused Crawler for Mining Hate and Extremism Promoting Users, Videos and Communities on YouTube”,2014 on “Best –first search and shark search”.


[9] Clementine Gritti1, Willy Susilo1, Thomas Plantard1, Kaitai Liang2, and Duncan S. Wong3 “Empowering Personal Health Records with Cloud Computing”