Topic-Specific YouTube Crawling to Detect Online Radicalization

Swati Agarwal[1] and Ashish Sureka[2]

[1] Indraprastha Institute of Information Technology-Delhi (IIIT-D), India
swatia@iiitd.ac.in
[2] Software Analytics Research Lab (SARL), India
ashish@iiitd.ac.in

Abstract. Online video sharing platforms such as YouTube contains several videos and users promoting hate and extremism. Due to low barrier to publication and anonymity, YouTube is misused as a platform by some users and communities to post negative videos disseminating hatred against a particular religion, country or person. We formulate the problem of identification of such malicious videos as a search problem and present a focused-crawler based approach consisting of various components performing several tasks: search strategy or algorithm, node similarity computation metric, learning from exemplary profiles serving as training data, stopping criterion, node classifier and queue manager. We implement two versions of the focused crawler: best-first search and shark search. We conduct a series of experiments by varying the seed, number of n-grams in the language model based comparer, similarity threshold for the classifier and present the results of the experiments using standard Information Retrieval metrics such as precision, recall and F-measure. The accuracy of the proposed solution on the sample dataset is 69% and 74% for the best-first and shark search respectively. We perform characterization study (by manual and visual inspection) of the anti-India hate and extremism promoting videos retrieved by the focused crawler based on terms present in the title of the videos, YouTube category, average length of videos, content focus and target audience. We present the result of applying Social Network Analysis based measures to extract communities and identify core and influential users.

Keywords: Mining User Generated Content, Social Media Analytics, Information Retrieval, Focused Crawler, Social Network Analysis, Hate and Extremism Detection, Video Sharing Website, Online Radicalization.

1 Research Motivation and Aim

YouTube is a most popular video sharing website that allows users to watch and upload an unlimited number of videos. It also allows users to interact with each other by performing many social networking activities. According to YouTube statistics[1], over 6 billion hours of video are watched each month on YouTube.

[1] http://www.youtube.com/yt/press/statistics.html

W. Chu et al. (Eds.): DNIS 2015, LNCS 8999, pp. 133–151, 2015.

100 millions of people perform social activities every week and millions of new subscriptions are made every day. These subscriptions allow a user to connect to other users[2]. The high reachability 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 misuse YouTube in many ways by uploading malignant content that are offensive and illegal. For example, harassment and insulting videos [19], video spam [16], pornographic content [3], hate promoting [17] and copyright infringed videos [2].

Research shows that YouTube has become a convenient platform for many hate and extremist groups to share information and promote their ideologies. The reason because video is the most usable medium to share views with others [6]. Previous studies show that extremist groups put forth hateful speech, offensive comments and messages focusing their mission [11]. Social networking allows these users (uploading extremist videos, posting violent comments, subscribers of these channels) to facilitate recruitment, gradually reaching world wide viewers, connecting to other hate promoting groups, spreading extremist content and forming their communities sharing a common agenda [7] [20].

Online radicalization and extremism have a major impact on society that contributes to the crime against humanity[3]. The presence of such extremist content in large amount is a major concern for YouTube moderators (to uphold the reputation of the website), government and law enforcement agencies (identifying extremist content and user communities to stop such promotion in country). However, despite several community guidelines and administrative efforts made by YouTube, it has become a repository of large amounts of malicious and offensive videos [17]. Detecting such hate promoting videos and users is significant and technically challenging problem. 100 hours of videos are uploaded every minute, that makes YouTube a very dynamic website. Hence, locating such users by keyword based search is overwhelmingly impractical. The work presented in this paper is motivated by the need of a solution to combat and counter online radicalization. We frame our problem as 1) identifying such videos and users, promoting hate and extremism (Focus of this paper) on YouTube, 2) locating virtual and hidden communities of hate promoting users sharing a common goal or group mission and 3) identifying users with strong connections and playing central role in a community.

The research aim of the work presented in this paper is the following

1. To investigate the application of a focused crawler (best first search and shark search) based approach for retrieving YouTube user-profiles promoting hate and extremism. Our aim is to examine the effectiveness of two versions of the focused crawler (best-first search and shark search) and measure performance by varying experimental parameters such as the size of the n-gram, similarity threshold and seed.

[2] http://www.jeffbullas.com/2012/05/23/

[3] http://curiosity.discovery.com/question/how-hate-crime-impact-society

Table 1. Summary of Literature Survey of 14 Papers, Arranged in Reverse Chronological Order, Identifying Hate & Extremist Content on Various Platforms. VS= Video Sharing Websites, MB= Micro-Blogging, BL- Blogging, SN= Social Networking, DF= Discussion Forum, OW= Other Websites.

S.No.	Research Study	Platform	Objective & Analysis
1.	O'Callaghan et. al; 2013	MB	Analysis of extreme right activities on multiple platforms for community detection.
2.	I-Hsien Ting et. al; 2013	SN	Identifying extremist groups on Facebook using keywords and social network structure.
3.	G. Patil et. al; 2013	OW	Identifying and blocking terrorist websites using content analysis.
4.	M. GoodWin; 2013	MB, VS, SN	Analysis of various counter-jihad, Islam and Muslim communities on web 2.0.
5.	P. Wadhwa et. al; 2013	MB, VS, SN	Dynamic tracking of radical groups on web 2.0 by analyzing messages and post.
6.	H. Chen et. al ; 2012	DF, VS	Examine several dark web forums and videos used by terrorist & extremist groups.
7.	D. Denning et. al; 2012	VS, SN	An in-depth research on Social Media associated with jihad and counter terrorism.
8.	C. Logan, et. al ; 2012	BL, OW	Finding similarities between different extremist groups using thematic content analysis.
9.	S. Mahmood; 2012	MB, VS, SN	Comparing several defense mechanisms to detect terrorists on social network websites.
10.	J. Hawdon; 2012	MB, VS, SN	A statement about the effect of hate groups as hate-inspired violence on the web.
11.	O'Callaghan et. al; 2012	MB, VS, SN, OW	Activity and links analysis of extreme right groups (local and international) on Twitter.
12.	E. Erez; 2011	DF	Quantitative and qualitative assessments of the content of communications on forums.
13.	D. David et. al; 2011	MB, SN	Detecting criminal groups & most visible players using the keyword search & contacts.
14.	A. Sureka et. al.; 2010	VS	Locating hate promoting videos, users and their groups sharing a common agenda.

2. To investigate the effectiveness of contextual features such as the title of the videos uploaded, commented, shared, and favourited for computing the similarity between nodes in the focused crawler traversal. To examine the effectiveness of subscribers, featured channels and public contacts as links between nodes.
3. To conduct a case-study by defining a specific topic (anti-India) and perform an in-depth empirical analysis on real-world data from YouTube.
4. To conduct a characterization study of the anti-India hate and extremism promoting videos based on terms present in the title of the videos, YouTube category, average length of videos, content focus and target audience
5. To discover user communities and groups and apply Social Network Analysis (SNA) based measures (such as centrality) to identify core users.

2 Related Work and Research Contributions

In this section, we discuss closely related work to the study presented in this paper. We conduct a literature survey on the topic of hate and extremist content detection on Web 2.0. Table 1 shows a list of 14 papers in reverse chronological order. As shown in Table 1, we characterize the papers on the basis of the social media platform and the objective of analysis. Table 1 reveals that researchers have

conducted experiments on several social media platforms such as video-sharing website, micro-blogging websites, online discussion forums and social networking websites. Table 1 shows a diverse domain of study covered in existing literature: terrorism, extremist groups, anti-black communities, US domestic, middle eastern, jihad and anti-Islam.

1. I-Hsien Ting et. al. propose an architecture to discover hate groups on Facebook using text mining and social network analysis. Extracted features include keywords that are frequently used in groups [18].
2. M. Goodwin 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 [9].
3. A. Sureka et. al. 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 [17].
4. H. Chen et. al present a framework to identify extremist videos on YouTube. They extracted lexical, syntactic and content specific features from user generated data and applied different feature based classification techniques to classify videos [4] [6] [5] [8].
5. E. Reid et. al present a hyperlink study to discover US extremist groups and their online communities on various discussion forums and video sharing websites. They perform web crawling and text analysis on the web content in order to find the relevant websites [14].
6. A. Salem et. al propose a multimedia and content based analysis approach to detect jihadi extremist videos and the characteristics to identify the message given in the video [15].

In context to existing work, the study presented in this paper makes the following unique contributions (the study presented in this paper is an extension of our previous work [1]):

1. We present an application of focused or topical crawler based approach for locating hate and extremism promoting channels on YouTube. While there has been a lot of work in the area of topical crawling of web-pages, this paper presents the first study on adaptation of focused crawler framework (best-first search and shark-search) for navigating nodes and links on YouTube.
2. We conduct a series of experiments on real-world data downloaded from YouTube to demonstrate the effectiveness of the proposed solution approach by varying several algorithmic parameters such the size of n-gram for language modeling based statistical model, similarity threshold for the text classifier, starting point or seed for best-first search and shark search version of the algorithm.
3. We perform a characterization study of the anti-India hate and extremism promoting videos based on terms present in the title of the videos, YouTube category, average length of videos, content focus and target audience. We apply

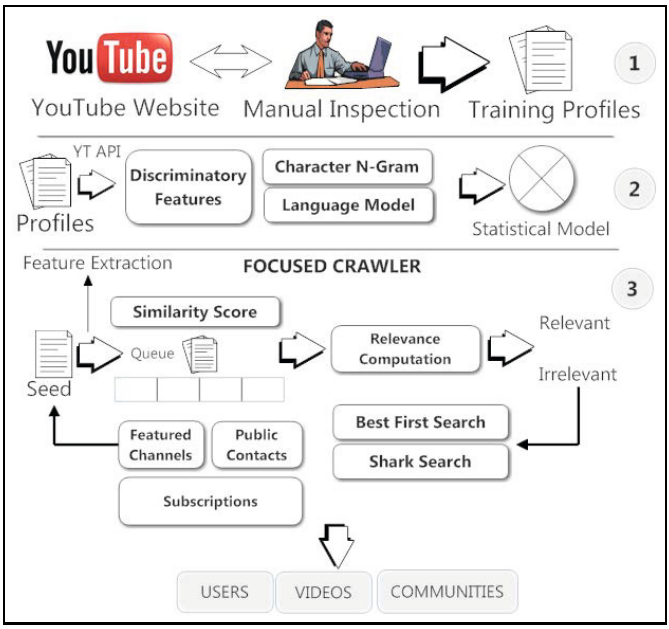


Fig. 1. A General Research Framework For Our Proposed Solution Approach

Social Network Analysis (SNA) based techniques on the retrieved user profiles and their connections obtained from the focused crawler traversal to understand presence of communities and central users.

3 Research Framework and Methodology

Figure 1 presents a general framework for the proposed solution approach. The proposed method is a multi-step process primarily consists of three phases, Training Profile Collection, Statistical Model Building and Focused Crawler cited as Phase 1, 2 and 3 respectively.

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, favourited & commented by the user and profile information) of these 35 channels using YouTube API[4]. In the training dataset, we observe several terms relevant to hate and extremism and divide them into 9 main categories shown in Table 2. We build a statistical model from these training profiles by applying character n-gram based language modeling approach.

We chose character-level analysis (low-level features) as it is language independent and does not require extensive language specific pre-processing. The other

[4] `https://developers.google.com/youtube/getting_started`

Table 2. Categorization of Sample Terms Occuring in Examplary Documents for Focused Crawler

Category	Terms
Important Dates	13th January, 26th January, 23rd March, 5th August, 14th August, 15th August, 21st September, 9th November, 3rd December, 25th December
Region	Hindustan, Pakistan, India, Kashmir, Bhindustan, Lahore, Afganistan, America, China, Turkey, Mumbai, Khalistan, Indo-Pak, US, Jammu & Kashmir, Agartala, Bangladesh, England, Israel, Karbala, Arabia, Argentine, Syria, Egyptian, Goa, Orrisa, Bihar, Canadian, Arab, Sindh, Balochistan, Punjab
Religion	Islam, Muslim, Hindu, Allah, Khuda, Quran, Maulana, Mosque, Kabba, Jihad, Azan, Jewish, Burka, Prophet, Religious, Koum, Islamic, Jews Christians, Apostates, Sikh, Buddhist, Hinduism, Muhajirs, Immigrant Muslims
People Name	Obama, Osama, Laden, Zaid Hamid, Zakir Naik, Parvez Musharraf, Mark Glenn, Jinnah, Saed Singh, Imran Khan, Nawaz Shareef, Quaid, Iqbal, Tahir Ashrafi, Emad Khalid, Yousuf Ali, Shaykh Feiz, Mustafa Kamal, Khalid Yaseen, Asma Jahangir, Chandragupt, Gandhi, Nehru, Pramod Mahajan
Negative Emotions	Horrible, Hate, Hatred, Murder, Cheating, Ice-Blood, Honour, Loathing, Humanity, Violence, Bloody, Blood, Revenge, Torture, Extremism, Humiliation, Abuse, Poverty, Fear, Scoundrel, Lies, Fraud, Friendship, Hesitation, Fake, Filthy, Discrimination
Communities	Laki Punjabi, CIA, ISI, Takmel-E-Pakistan, Brass Tacks, Azad Kashmir, Liberate Kashmir, Taliban, Aman Ki Asha, Flag Attack, Gang, IAF, Air Force, RAW, PMLN, NATO, TTP, Threek-E-Taliban, SWAT, WUP, PPP, Pakistani People Party, Operation Shudhi Karan, Aryavrat
Politics Terms	Conspiracy, Leader, Democracy, Inqalab, Awami, Strike, Khilafat, Against, Rights, Partition, Corruption, Media, Resolution, Objective, Rule, Party, League, Protest, Politician, Slogan, Division, Public, President, Secularism, Domestic, Congress, Election, Witnessed, Tribal, Rallies, persecuted, Youth
War Related Terms	LOC, Bomb, Blast, Attack, Holywar, Warfare, Tribute, Soldier, Jawan, Refugee, Enemies, Fighting, Patriot, Assassination, Expose, Propaganda, Army, Protocol, Security, Anthem, Threat, Nukes, Border, Shaheedi, Military, Zindabad, Hijab, Dirty War, Black Day, Terror, Mission, Operation, Jail, Prison, Open Fire, Destruction, Halaal, Grave, Sectarian, Genocide, Encounter, Ghadar, Strategy, Battle Field, Nation, Warning, Killing, Legendary, Campaign, Ghulami, Weapon, Qarz, Unsafe, Insulting, Defend, Accident, Judicial, Failure, Camp, Evil, Vision, Armed Forces, Agent, Martyred, Missing, Intentions, Defeat, Secret, Slap, Traitor, Reclaim, Tragedy, Shahadat, Accusing, TAKBEER, Terrorists, War On Terror, Crime, Bloodshed, Revolution, Constitution, Vandalism, Victorious, Violation, Graves, Torture, Slaughtered, Explodes, Struggle, War, Freedom, Jet Carrier, Police men, Slave, Honour killing
Others	Pig, Monkey, Faith, Ideology, Earthquake, Thunder, Uneducated, Awareness, Debate, Foreign, Leaked, Press, Affair, Economic, Destiny, Flood, Endgame, Rebuttal, Documentary, Respect, Argue, Patrol, Scandal, Survival, Rapist, Rape, Ideological, Geographical, Sections, Sects, Government, Interview

advantage of character n-gram based approach is that it can capture sub-word and super-word features and is suitable for noisy text found in social media. The paper by Peng et al. lists the advantages of character-level n-gram language models for language independent text categorization tasks [12]. In phase 3, we build a focused crawler (best first search and shark 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. We find the extent of textual similarity between these metadata and training data by using statistical model (build in phase 2) and LingPipe API[5]. 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 predicted threshold. If a channel is relevant, then we further extend it's frontiers (links to other YouTube

[5] `http://alias-i.com/lingpipe/index.html`

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 library[6]. 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. We perform social network analysis on the output graph to locate hidden communities of hate promoting users.

3.1 Solution Implementation

In this section, we present the methodology and solution implementation details for the design and architecture articulated in the previous section. In focused crawler we first classify a seed input as relevant or irrelevant which further leads to more relevant channels. In proposed method we use focused crawler for two different graph traversing algorithms i.e. Best First Search (BFS) Algorithm and Shark Search Algorithm (SSA). Algorithm 1 and Algorithm 2 describe the focused crawlers we develop to locate a group of connected hate and extremist channels on YouTube. The result of both algorithms is turned out to be a directed cyclic graph where each node represents a user channel and an edge represents a link between two users. The goal of BFS and SSA is to first classify a channel to be relevant (positive class) or irrelevant (negative class) and then exploring the frontier channels of a relevant user (in case of BFS) and both users (in case of SSA).

Inputs to these algorithms are a seed (a positive class user) U, width of graph w i.e. maximum number of children of a node, size of graph s i.e. maximum number of nodes in graph, threshold th for classification, n-gram value Ng for similarity computation, and a lexicon of 35 positive class channels U_p. Table 4 shows a list of all seed inputs we have used for different iterations. We compare each training profile with all profiles and compute their similarity score for each mode. We take an average of these 35 scores and compute the threshold values. Both algorithms are different in their approach explained in following subsections:

Focused Crawler- Best First Search. The proposed method (Algorithm 1) follows the standard best first traversing to explore relevant user to seed input. Best-First Search examines a node in the graph and finds the most promising node among it's children to be traversed next [13]. This priority of nodes (users) is decided based upon the extent of similarity with the training profiles. A user with the similarity score above a specified threshold is said to be relevant and allowed to be extended further. If a node is relevant and has the highest priority (similarity score) among all relevant nodes then we extend it first and explore it's links and discard irrelevant nodes. We process each node only once and if a node appears again then we only include the connecting edge in the graph.

Steps 1 and 2 extract all contextual features for 35 training profiles using Algorithm 3 and build a training data set. Algorithm SSA is a recursive function

[6] `http://jsoup.org/apidocs/`

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 ∈ 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
11 |  |  for all Ug ∈ Ugraph do
12 |  |  |  fr = Extract_Frontiers(Ug)
13 |  |  |  Hashmap Ucrawler.add(fr)
   |  |  end
14 |  |  for all Ufr ∈ Ucrawler do
15 |  |  |  BFS(Ufr)
   |  |  end
   |  end
```

which takes U as a seed input. Steps 4 and 5 extract all features for seed user U and compute it's similarity score with training profiles using character n-gram and language modeling (using LingPipe API). Steps 6 to 8 represent the classification procedure and labeling of users as relevant or irrelevant depending upon the threshold measures.

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 16 extracts frontiers of a user channel using Algorithm 4. Steps 18 and 19 repeat steps 3 to 15 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.

Focused Crawler- Shark Search. We propose a focused crawler for Shark Search Algorithm (Algorithm 2), an adaptive version of the same algorithm introduced in M. Hersovici et. al. [10]. 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.

Steps 1 to 5 are similar to Best First Search (Algorithm 1). Steps 6 to 9 check if the user is a child of irrelevant node then it computes an inherited score for the user by multiplying the original score by a decay factor d. If a node has appeared before and has not been extended further then we update it's similarity score by the maximum value of old and new inherited score. Steps 10 to 12 represent the

Algorithm 2: Focused Crawler- Shark Search

```
   Data: Seed User U, Width of Graph w, Size of Graph s, Threshold th, N-gram Ng, Positive Class Channels Up, Decay Factor d
   Result: A connected directed cyclic graph, Nodes=User u
 1 for all u ∈ Up do
 2 |   D.add(ExtractFeatures(u))
   end
   Algorithm SSA(U)
 3 |   while graphsize < s do
 4 |   |   userfeeds Uf ←ExtractFeatures(U)
 5 |   |   score score ←LanguageModeling(D, Uf, Ng)
 6 |   |   if (U is a child of Irrelevant node) then
 7 |   |   |   score ← score * d
   |   |   end
 8 |   |   if (U has appeared before) then
 9 |   |   |   score ← max(new_score, old_score)
   |   |   end
10 |   |   if (score <th) then
11 |   |   |   U.newclass ←Irrelevant
   |   |   else
12 |   |   |   U.newclass ←Relevant
   |   |   end
13 |   |   Hashmap Usorted.InsertionSort(U, score)
   |   |   for i ← 1 to w do
14 |   |   |   Hashmap Ugraph.add(Usorted(i))
   |   |   end
15 |   |   for all Ug ∈ Ugraph do
16 |   |   |   fr = Extract_Frontiers(Ug)
17 |   |   |   Hashmap Ucrawler.add(fr)
   |   |   end
18 |   |   for all Ufr ∈ Ucrawler do
19 |   |   |   SSA(Ufr)
   |   |   end
   |   end
```

classification procedure and labeling of users as relevant or irrelevant similar to Algorithm 1.

The SSA method also uses non-binary priority values same as similarity score of users. Steps 13 and 14 make a list of top w (maximum number of children, a node can have) users (could be relevant or irrelevant unlike BFS) based upon their similarity score, sorted in a decreasing order. Steps 15 to 19 extract frontiers of a user channel using Algorithm 4 and repeats steps 3 to 19 for each linked user.

Features Extraction. In Algorithm 3, we retrieve contextual metadata of a YouTube user channel using YouTube API. Step 1 extracts the profile summary of the user. Steps 2 to 5 extract the titles of videos uploaded, commented, shared and favourited by given user U. The result of the algorithm is a text file containing all the video titles and user profile information.

Algorithm 3: Features Extraction for a YouTube User

```
   Data: User u
   Result: User Activity Feeds and Profile Information
   Algorithm ExtractFeatures(U)
1  |   uProfile ←u.getSummary()
2  |   uUploads ←u.getUploadedVideo()
3  |   uCommented ←u.getCommentedVideo()
4  |   uShared ←u.getSharedVideo()
5  |   ufavorited ←u.getFavoritedVideo()
```

Algorithm 4: Frontier Extraction for a YouTube User

```
Data: User u
Result: Frontiers of a channel
Algorithm Extract_Frontiers(U)
1   u_subs ←u.getSubscribers()
2   u_fc ←u.getFeaturedChannels()
3   u_con ←u.getFriends()
```

Table 3. List of Few Users Ids of Hate and Extremism Promoting Videos Being Used As Exemplary Documents For Training A Text Classifier

AabeKosar	BTghazwa	haider2026	IndianVictim
Ahmad12791	charbi88	issabln2011	kashafsha
amiruddinmughal	GobletG	GreaterPakistan	khawajak
azadkashmiriboy	hijazna	HinduismIslam	junihashmi
BrassTacksOfficial	netdarvin	IndiaEternal	GreenEye1947
PakistanKaKhudaHafiz	p4pathanp4pakistan	sabeqoonwaawaloon	TAKMEELEPAKISTAN

Table 4. Name of 10 Seed Inputs Used for BFS and SSA- Row-wise Ordered

TheGreaterPakistan	BTghazwa	GreaterPakistan	PakistanRoxxx	PakistanKaKhudaHafiz
BrassTacksOfficial	haider2026	hiddenpakistani	PakistanHeaven	MujheHayHukmeAzan

Frontiers Extraction. In Algorithm 4, we extract all external links of a YouTube channel to other YouTube channels. These links could be the subscribers, featured channels (suggestions by user) and public contacts (friends). YouTube API does not allow users to retrieve the contacts of other users which is why we use jsoup HTML parser library to fetch all frontiers and public contacts list. This algorithm returns a vector of all channels user U is linked with and we make sure that there is no redundant channel in the list.

4 Empirical Analysis and Performance Evaluation

In this section we present the characterization of hate and extremist videos. We demonstrate the experiments and analysis set up, performance results and the effectiveness of our proposed solution approach.

4.1 Experimental Dataset

Training Dataset. A focused crawler needs to classify if a given web-page is relevant or not with respect to a topic. The crawler requires exemplary documents or training examples to learn the specific characteristics and properties of documents in the training dataset. A statistical model (text classifier) needs to be built from a collection of documents pertaining to a predefined topic. Table 3 shows a list of few user ids (channel names on YouTube) used as a training profiles. These user ids consists of 612 videos and hence the training is performed on 612 videos. We obtain the training dataset by manually searching (keyword

Table 5. Results of Focused Crawler for 6 Different Seeds. Modes Represent 6 Different Thresholds (Th) & N-gram (Ng) Pairs. A: Th=-2.0, Ng=3, B: Th=-2.5, Ng=3, C: Th=-3.0, Ng=3, D: Th=-2.0, Ng=5, E: Th=-2.5, Ng=5, F: Th=-3.0, Ng=5.

(a) Focused Crawler- Best First Search

Seed	Seed 1						Seed 2						Seed 3					
Mode	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F
Relevant	26	19	58	21	56	60	39	57	30	26	64	67	1	1	1	1	1	1
Irrelevant	23	2	9	3	11	7	11	1	4	5	1	1	0	0	0	0	0	0
Processed	119	448	239	134	159	145	119	448	239	134	312	263	1	1	1	1	1	1
Graph	23	19	25	21	26	26	23	24	25	22	26	26	1	1	1	1	1	1
Minimum	-2.13	-2.7	-6.83	-6.3	-4.75	-4.75	-9.43	-9.43	-6.83	-0.84	-8.43	-8.43	-1.78	-1.78	-1.78	-1.01	-1.01	-1.01
Maximum	-2.13	-0.97	-0.52	-0.48	-0.48	-0.48	-0.8	-0.8	-0.52	-0.46	-0.46	-0.46	-1.78	-1.78	-1.78	-1.01	-1.01	-1.01
Median	-2.13	-1.73	-1.91	-1.23	-1.23	-1.23	-2.08	-1.85	-1.72	-1.7	-1.7	-1.7	-1.78	-1.78	-1.78	-1.01	-1.01	-1.01
Quartile 1	-2.13	-2.13	-2.21	-1.39	-1.78	-1.78	-2.26	-2.26	-2.19	-2.16	-2.08	-2.1	-1.78	-1.78	-1.78	-1.01	-1.01	-1.01
Quartile 3	-2.13	-1.3	-1.54	-0.87	-0.87	-0.87	-1.65	-1.58	-1.5	-1.2	-1.2	-1.19	-1.78	-1.78	-1.78	-1.01	-1.01	-1.01

Seed	Seed 6						Seed 7						Seed 8					
Mode	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F
Relevant	5	34	27	23	32	32	0	28	25	21	31	31	1	1	1	1	1	1
Irrelevant	2	4	10	11	6	6	1	5	1	4	2	2	0	0	0	0	0	0
Processed	20	313	290	258	332	332	0	212	318	256	252	274	1	1	1	1	1	1
Graph	5	22	25	22	25	25	0	22	25	21	26	26	1	1	1	1	1	1
Minimum	-2.25	-2.87	-6.83	-6.3	-2.38	-2.38	-2.04	-0.77	-6.83	-6.3	-4.75	-4.75	-1.87	-1.87	-1.87	-1.72	-1.72	-1.72
Maximum	-1.16	-0.97	-0.52	-0.46	-0.46	-0.46	-2.04	-0.97	-0.52	-0.46	-0.46	-0.46	-1.87	-1.87	-1.87	-1.72	-1.72	-1.72
Median	-1.6	-1.68	-1.78	-1.22	-1.29	-1.29	-2.04	-1.77	-1.91	-1.23	-1.33	-1.33	-1.87	-1.87	-1.87	-1.72	-1.72	-1.72
Quartile 1	-1.94	-2.13	-2.21	-1.59	-1.91	-1.91	-2.04	-2.13	-2.24	-1.79	-2.05	-2.05	-1.87	-1.87	-1.87	-1.72	-1.72	-1.72
Quartile 3	-1.34	-1.3	-1.46	-0.82	-0.46	-0.46	-2.04	-1.31	-0.52	-0.89	-1.12	-1.12	-1.87	-1.87	-1.87	-1.72	-1.72	-1.72

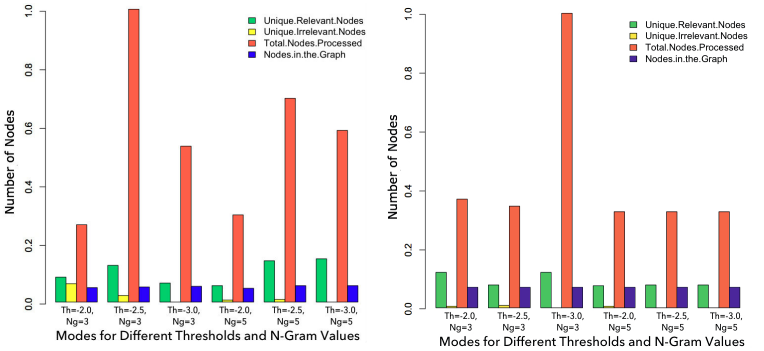
(a) Focused Crawler- Best First Search

Seed	Seed 1						Seed 2						Seed 3					
Mode	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F
Relevant	37	34	27	56	29	27	45	29	45	28	29	29	1	1	1	1	1	1
Irrelevant	6	2	2	3	3	2	2	3	0	2	0	0	0	0	0	0	0	0
Processed	198	167	123	177	110	110	138	129	374	122	122	122	1	1	1	1	1	1
Graph	26	26	26	26	26	26	25	26	26	26	26	26	1	1	1	1	1	1
Minimum	-13.9	-13.9	-13.9	-13.2	-132	-132	-2.25	-2.70	-2.70	-2.31	-2.42	-2.42	-1.78	-1.78	-1.78	-1.01	-1.01	-1.01
Maximum	-0.11	-0.12	-0.17	-0.07	-0.15	-0.15	-0.11	-0.13	-1.05	-0.10	-0.46	-0.46	-1.78	-1.78	-1.78	-1.01	-1.01	-1.01
Median	-0.50	-1.46	-1.70	-1.20	-1.21	-1.41	-1.62	-1.47	-1.74	-0.81	-1.18	-1.18	-1.78	-1.78	-1.78	-1.01	-1.01	-1.01
Quartile 1	-1.47	-2.04	-2.26	-1.57	-1.91	-1.97	-1.79	-1.94	-2.30	-1.22	-1.96	-1.96	-1.78	-1.78	-1.78	-1.01	-1.01	-1.01
Quartile 3	-0.21	-0.25	-1.39	-0.85	-0.71	-0.92	-1.26	-1.08	-1.30	-0.23	-0.86	-0.86	-1.78	-1.78	-1.78	-1.01	-1.01	-1.01

Seed	Seed 6						Seed 7						Seed 8					
Mode	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F
Relevant	23	35	27	37	23	23	22	36	26	36	23	23	1	1	1	1	1	1
Irrelevant	4	3	0	0	2	2	2	3	0	3	2	2	0	0	0	0	0	0
Processed	158	169	88	131	80	80	242	213	65	107	54	54	1	1	1	1	1	1
Graph	25	25	25	25	25	25	25	25	25	25	21	21	1	1	1	1	1	1
Minimum	-2.32	-2.70	-2.70	-2.31	-3.62	-3.62	-2.32	-2.70	-2.70	-2.31	-3.62	-3.62	-1.87	-1.87	-1.87	-1.72	-1.72	-1.72
Maximum	-0.05	-0.13	-0.33	-0.07	-0.46	-0.46	-0.05	-0.12	-0.33	-0.07	-0.17	-0.17	-1.87	-1.87	-1.87	-1.72	-1.72	-1.72
Median	-0.23	-1.63	-1.63	-0.87	-1.23	-1.23	-0.20	-1.60	-1.63	-0.97	-1.23	-1.23	-1.87	-1.87	-1.87	-1.72	-1.72	-1.72
Quartile 1	-1.40	-2.12	-2.05	-1.32	-2.05	-2.05	-0.26	-2.25	-1.97	-1.33	-2.05	-2.05	-1.87	-1.87	-1.87	-1.72	-1.72	-1.72
Quartile 3	-0.16	-1.12	-1.60	-0.39	-0.93	-0.93	-0.16	-1.16	-1.63	-0.45	-0.78	-0.78	-1.87	-1.87	-1.87	-1.72	-1.72	-1.72

based) for anti-India hate and extremism promoting channels using YouTube search and traversing related video links (using the heuristic that videos on similar topic will be connected as relevant on YouTube). The training dataset profile consists of profile information of users and the title of videos uploaded, favorited, shared and commented by the user. We believe the title of such videos reflects user interests and can be used for building a predictive model.

Test Dataset. We select 10 random positive class (hate and extremist) channels for creating test dataset. Each user works as a seed input to the focused crawler. Table 4 shows the list of all 10 seeds we select for our experiments. To evaluate the effectiveness of our solution approach we execute our focused crawler sixty times for both Shark Search and Best First Search. Here we use 10 different seeds, 3 different threshold values and 2 different n-gram values for similarity computation. We make 6 pairs of threshold and n-gram values calling them as six different "Modes". For both approaches (BFS and SSA), we run our focused crawler 60 times for 10 seeds and each seed for all 6 modes.



(a) Best First Search (b) Shark Search

Fig. 2. Illustrating The Variance Between Number of Unique Relevant Nodes, Unique Irrelevant Nodes, Nodes Present in The Graph and Total Number of Nodes Processed for Six Different Modes of Seed 2

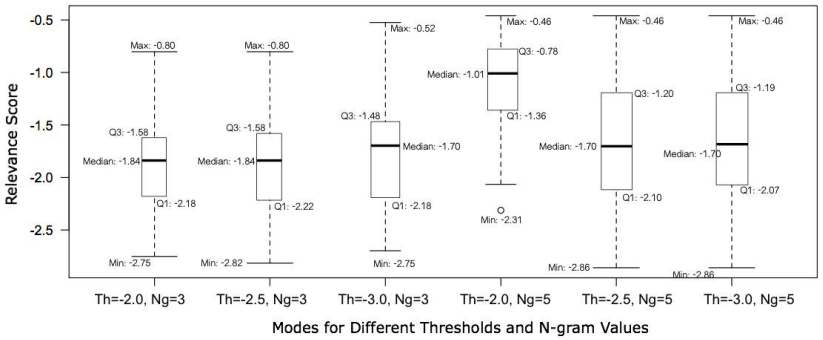


Fig. 3. Box-Plot And Descriptive Statistics For Six Different Configurations Of Best First Search Crawler

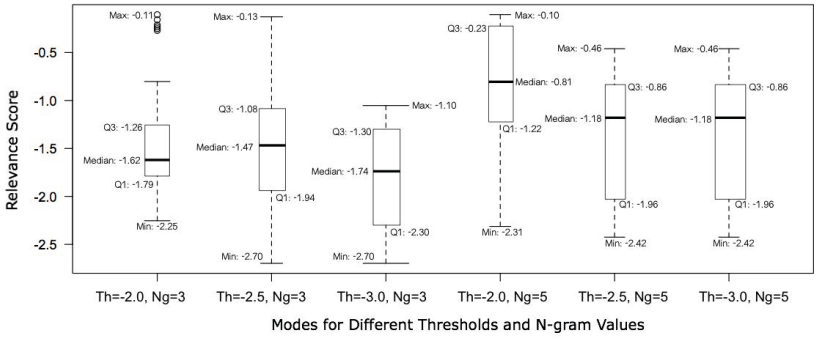


Fig. 4. Box-Plot and Descriptive Statistics for Six Different Configurations of Shark Search Crawler

4.2 Experimental Results

Focused Crawler Results. As mentioned above, we execute our focused crawler 60 times for both BFS and SSA. Table 5 (a) and (b) present the complete picture of users statistics based upon their similarity scores in each iteration. Table 5 (a) and (b) show the number of unique relevant users, unique irrelevant users, total number of users present in the output graph and the total number of users processed during execution of BFS and SSA focused crawlers respectively. Table 5 also shows the summary of similarity scores (minimum, maximum, median, 1^{st} quartile and 3^{rd} quartile) of all users. Table 5 reveals that the number of relevant and irrelevant users vary for different threshold and n-gram pairs. In Table 5 we notice that for both BFS and SSA, five-gram performs better than tri-gram. And for five-gram we achieve maximum number of relevant users in mode F (threshold= -3.0, n-gram= 5). These statistics show that the number of relevant and irrelevant nodes vary for different seeds. For example, for seed 3 and 8 we have only one relevant node. Despite being positive class channels these users have no links to other hate and extremist users on YouTube. Table 5 (a) and (b) reveal the difference in BFS and SSA performance for same seed. For seed 7, 9 and 10, we have an empty graph for BFS while in SSA we have 25 connected users for mode A. And similarly for other modes SSA has more number of relevant users in comparison to BFS.

Figure 2(a) and 2(b) illustrate the variance in number of nodes (shown on Y-axis) for different modes (shown on X-axis) for one seed. Where each node represents a YouTube user. Figure 2(b) depicts that for each mode number of irrelevant nodes for SSA are negligible in comparison to BFS. We also notice that for Seed 2, the graph size is almost similar in both BFS and SSA approach. In BFS we extract frontiers of only relevant nodes unlike SSA. Therefore, for BFS, we see a radical change in number of processed nodes for each mode. For SSA the number of unique relevant nodes as well as the number of processed nodes are similar for all modes except mode C. Figure 3 and 4 show the variance in the statistics of similarity or relevance score (shown on Y-axis) for different modes (shown on the x-axis). These statistics are measured for one seed used for both BFS and SSA approaches and same configuration of threshold and n-gram values. In Figure 3 we see that the first quartile for mode A is below the threshold value and it is smaller than third quartile unlike in Figure 4. It is an evidence that for BFS the number of relevant nodes are lesser in comparison to SSA. In SSA approach we are able to find users which are more relevant (shown as outliers) to training profiles. Figure 3 and 4 show that for modes E and F (Th=-2.5, Ng=5 and Th=3, Ng=5 respectively) all users are classified at relevant.

We asked 3 graduate students of our department to validate our results and they manually annotated each user. Based upon the validation we evaluate the accuracy of our classifier by comparing the predicted class against the actual class of each user channel. Table 6(a) shows the confusion matrix for binary classification performed during Best First Search approach. Given the input of 10 seed users and 6 modes (pair of threshold and n-gram values) we get different

Table 6. Confusion Matrix for Focused Crawlers

(a) Best First Search

		Predicted	
		Relevant	**Irrelevant**
Actual	**Relevant**	991	295
	Irrelevant	55	29

(b) Shark Search Algorithm

		Predicted	
		Relevant	**Irrelevant**
Actual	**Relevant**	921	314
	Irrelevant	125	67

Table 7. Accuracy Results for Focused Crawler- Best First Search and Shark Search. TPR= True Positive Rate, FPR= False Positive Rate, PPV= Positive Predictive Value, NPV= Negative Predicted Value.

	TPR	**TNR**	**PPV**	**NPV**	**F1-Score**	**Accuracy**
BFS	0.75	0.35	0.88	0.18	0.81	0.69
SSA	0.77	0.35	0.95	0.09	0.85	0.74

Table 8. Illustrating The Network Level Measurements for Focused Crawlers- Best First Search (Left) and Shark Search Algorithm (SSA). NN= Number of Nodes, NE= Number of Edges, SL= Number of Self Loops, Dia= Network Diameter, AD= Average Density, ACC= Average Clustering Coefficient, IBC= In- Betweenness Centrality, CC= In- Closeness Centrality, #W/SCC= Number of Weak/Strong Connected Components.

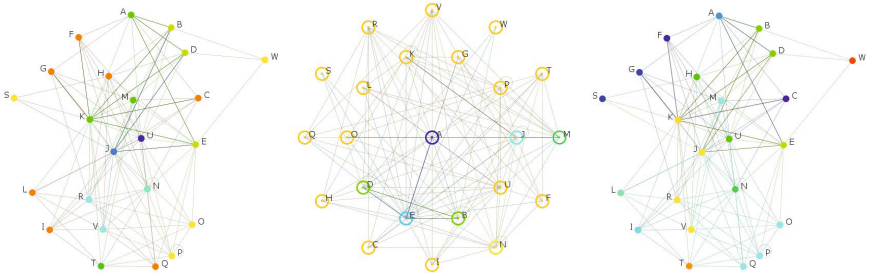
	NN	**NE**	**SL**	**Dia**	**AD**	**ACC**	**IBC**	**ICC**	**#WCC**	**#SCC**
BFS	23	119	3	4	0.225	0.388	0.046	0.356	1	7
SSA	24	137	8	3	0.238	0.788	0.009	0.320	1	16

number of connected users in each iteration. To measure the accuracy of our proposed approach we collect results of all 60 iterations and classify 1046 (921 + 125) users as relevant and 381 (314 + 67) as irrelevant users. There is a misclassification of 25.42% and 65.10% in predicting the relevant and irrelevant users respectively. Table 6(b) shows the confusion matrix for binary classification during Shark Search approach. Given the input of 10 seed users and 6 n-gram & threshold pairs, it classifies 1046 (991 + 55) users as relevant and 324 (295 + 29) as irrelevant users. There is a misclassification of 22.93% and 65.47% in predicting the relevant and irrelevant users respectively. This misclassification occurs because of the noisy data such as lack of information, non-english text and misleading information.

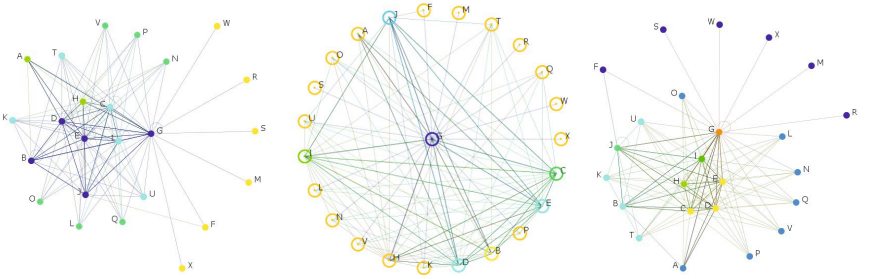
Table 7 shows the accuracy results (precision i.e. PPV, recall i.e. TPR, NPV, TNR, f1-score and accuracy) of focused crawler for both Best First Search and Shark Search approaches. Table 7 reveals that overall SSA approach (accuracy of 74%) performs better than BFS approach (accuracy 69%). Precision and accuracy of SSA are much higher than BFS and similarly recall and f1- score are reasonably higher for SSA.

Social Network Analysis. We perform social network analysis on the output graph of focused crawler, where each node represents a YouTube user channel and each edge represent a relation (friend, subscriber and featured channel) between two users. Table 8 illustrate the network level measurements we perform on the output graphs of BFS and SSA focused crawlers. These values have been computed for seed 2 in mode B (configuration of threshold=-2.5 and n-gram=3). In Table 8 we notice that in SSA approach users are strongly connected in comparison to BFS approach because the average density of network graph is more in SSA approach. Network diameter shows that in SSA each user is reachable in maximum 3 hops while in BFS it takes 4 hops. In SSA, we have more number of connected components than BFS, which helps to locate more communities. Here we see, that SSA has higher clustering coefficient which results into a cluster of highly relevant users.

Figures 5(a) and 5(b) (generated using ORA[7]) show three different representations of network graph, outputs for BFS and SSA focused crawler respectively (seed 2 and mode B- threshold=-2.5, n-gram=3). Graph in the left shows a directed connected cyclic graph. Colors of nodes represent the different in-degree of users and the width of an edge is scaled based upon the number of links between



(a) Best First Search Approach



(b) Shark Search Approach

Fig. 5. Community (Left), Betweenness Centrality (Middle) and Cluster (Right) Graph Representation for Best First Search (Top) Shark Search Crawler (Bottom) With Configuration: Th=-2.5, Ng=3 and Seed 2 (Node 'A')

[7] http://www.casos.cs.cmu.edu/projects/ora/

two users. In community graphs we see that for BFS all nodes are connected to each other unlike in SSA a few nodes are connected to only one user. Despite the existence of these nodes we find many strongly connected components in SSA which is very less in BFS because all nodes are equally connected. Graphs in the middle of Figures 5(a) and 5(b) are different representation of the output graph based upon the betweenness centrality. Node in the center has the highest centrality among all users and connected to all users of outer shells. In Figures 5(a) and 5(b), graphs in the right are the cluster representation of network. As we see in Table 8, the average clustering coefficient of network in SSA approach is very large in comparison to BFS. Similarly in the Figure 5(a) we see that in BFS approach network has 13 clusters, where total number of nodes is 23. Among these 13 clusters, 6 clusters have only one user node which shows the lack of similarity among users. In Figure 5(b) the cluster representation of network graph (right most graph) has 7 clusters for 24 nodes. where only 2 clusters are formed with one user. In this graph, each cluster shows the level of connectivity to other users. We see the existence of three strong communities made by nodes C, D, E, H, I and B, C, D, E, G, H, I, J and A, D, E, G, where G is the center of all communities and connected to all users.

Manual Analysis of Videos. We perform a manual analysis on YouTube and collect 274 hate and extremist videos uploaded by 35 unique users. We perform a characterization on these 274 and divide them into 5 different sets, shown in Table 9. We categorize these videos based upon three main parameters: 1) focus of the content shown in the videos, 2) targeted audience of the users uploading these videos and 3) the keywords presented in the title & description and used or spoken in the video. We also perform a characterization of these videos based upon the content shown in the video. Table 9 reveals that the average duration of these videos is from 3 minutes to 45 minutes. We observe that the 43 of total videos were small clips showing women and children harassment in India and Pakistan. For example, child labor, prostitution, slave. We find that majority of videos focus on islam promotion. These videos are very large in duration and defined as education videos on YouTube. Table 9 also shows that majority of videos fall under news and politics category and very few of them are uploaded for entertainment purpose. These videos target those audience who are affected by the incidents shown in the videos.

For example, 1947 partition, liberate Kashmir and hate speech videos against Pakistan and India targeting the haters of respective nations. The keywords shown in the Table 9 are the clear evidence of these videos to be hate promoting. We notice that all these videos are not just public recording, but users have used more creative ways to present their messages in front of their audiences. We divide these 274 videos into 12 categories based upon the type of content shown in the video. Table 10 shows that now users have used animation, cartoon, drawings, group discussions and textual messages in their videos to promote hate and extremism. These videos leave a negative impact on the audience and provoke them to write hateful comments.

Table 9. Categorisation of Videos Based Upon Keywords in Video Content & Title and Target Domain of the Uploader, #VD= Number of Videos, YT Category= Youtube Category, Avg Len= Average Duration of Video (in seconds)

#VD	YT Category	Avg Len	Content Focus	Target Audience	Keywords
43	News & Non-Profit	151.68	Honor Killing, Harassment	Women, Refugee People, Child	Honour Killing, Child Marriage, Rape, Responsibility, Protest, Women, Asylum, Arrested, Security, Safety, Refugee, Minor, Brutally, Beating, Exploit, Kidnap, Prostitute, Slave, Indian Police, Delhi, Child Labour.
93	News, Auto, Vehicle, Politics & Education	2526.16	Islam Promotion	Jewish And Muslim People	Vandalism, Jews, Christians, Apostates, Country, Shakyh Abu Hamza, Speech, Hate, Muslim, Fatah Domestic, Leader, Destroy, Killing, Rape, Taliban, Bombs, Battle, Courage, Allah, Islam, Courage, Principle, Politics, Belief, Macca, Trouble, Money, Hatred, Shaheed, Afganistan, Enemies Of Islam, Shame, Rape, Women, Kids, Bad, Women Rights, Debate, Zaid Hamid, Armed Force, Jinnah
25	News, Politics & Education	1225.56	Liberate Kashmir	Kashmiri People	Muslim, Army, Military, 1947, Partition, Azad Kashmir, Liberate Kashmir, Pakistan, India, Killing, Murder, Border, Fighting, Democracy, Martyr, Torture.
83	News & Politics	349.28	Anti-Muslims	Pakistan Haters	Kashmir, Jihad, Pakistan, India, Quran, Muslim, Hindu, Qatil, Zakir Naik, Hate Speech, Masjid, Pandit, Defense, Madarsa, Tribute, Bharat, America, Attack, Napak, Holy, Kabba, Prophet, Strike, Truth, Holy War, Jihad, Al-Queda, Blast, Killed, Enemies Of India, Leftists,Separatists, Maoists, Propaganda, Kasab.
30	Entertainment, Travel, News & Politics	319.61	Anti-India	India Haters	Kashmir, Poverty, Mumbai, Liberate, Hindu, Beggars, Human, Untouchable, Pundit, Casteism, Fraud, Extremism, Attack, Mob, Killed, Anti-Muslim, Anti-Pakistani, Hatred, Masks, Freedom.

Table 10. Categorization of Hate & Extremism Videos Based Upon the Content Shown in the Video

Pictures With Background Music, Animated Videos
Speech, News Segments, Drawing, Interviews , Group Discussion
Lectures, Cartoon And Comics, Debate, Recorded Videos, Textual Messages

5 Conclusions

We present a focused-crawler based approach for identification of hate and extremism promoting videos on YouTube. The accuracy for BFS and SSA versions of the algorithm is 0.69 and 0.74 respectively. Experimental results reveal higher precision, recall and accuracy for shark-search approach in comparison to best-first search. We conduct a series of experiments by varying various algorithmic parameters such as the similarity threshold for the language modeling based text classifier and n-grams. We conclude that by performing social network analysis on network graphs, we are able to locate hidden communities. We identify the users who play major roles in the communities and have highest centrality among all. We reveal the communities by dividing the network graph into clusters formed by similar users. In SSA we find more strongly connected components (16) and communities in comparison to BFS (7).

We perform a characterization on the content and contextual information of several hate promoting videos. The analysis reveals that hate promoting users upload videos targeting some specific audiences. Majority of videos are very large in the duration (3 to 45 minutes). Keywords present in the contextual information and video content are the evidence of these videos doing hate promotion among their viewers.

References

1. Agarwal, S., Sureka, A.: A focused crawler for mining hate and extremism promoting videos on youtube. In: Proceedings of the 25th ACM Conference on Hypertext and Social Media, HT 2014, pp. 294–296. ACM, New York (2014), http://doi.acm.org/10.1145/2631775.2631776
2. Agrawal, S., Sureka, A.: Copyright infringement detection of music videos on youtube by mining video and uploader meta-data. In: Bhatnagar, V., Srinivasa, S. (eds.) BDA 2013. LNCS, vol. 8302, pp. 48–67. Springer, Heidelberg (2013), http://dx.doi.org/10.1007/978-3-319-03689-2_4
3. Chaudhary, V., Sureka, A.: Contextual feature based one-class classifier approach for detecting video response spam on youtube. In: 2013 Eleventh Annual International Conference on Privacy, Security and Trust (PST), pp. 195–204 (2013)
4. Chen, H.: Extremist youtube videos. In: Dark Web. Integrated Series in Information Systems, vol. 30, pp. 295–318. Springer, New York (2012), http://dx.doi.org/10.1007/978-1-4614-1557-2_15
5. Chen, H., Denning, D., Roberts, N., Larson, C.A., Yu, X., Huang, C.-N.: Chapter 1 - revealing the hidden world of the dark web: Social media forums and videos. In: Yang, C., Mao, W., Zheng, X., Wang, H. (eds.) Intelligent Systems for Security Informatics, p. 1. Academic Press, Boston (2013), http://www.sciencedirect.com/science/article/pii/B978012404702000001X
6. Chen, H., Denning, D., Roberts, N., Larson, C.A., Yu, X., Huang, C.: The dark web forum portal: From multi-lingual to video. In: ISI, pp. 7–14. IEEE (2011), http://dblp.uni-trier.de/db/conf/isi/isi2011.html#ChenDRLYH11

7. Conway, M., McInerney, L.: Jihadi video and auto-radicalisation: Evidence from an exploratory youtube study. In: Ortiz-Arroyo, D., Larsen, H.L., Zeng, D.D., Hicks, D., Wagner, G. (eds.) EuroIsI 2008. LNCS, vol. 5376, pp. 108–118. Springer, Heidelberg (2008), http://dx.doi.org/10.1007/978-3-540-89900-6_13
8. Fu, T., Chen, H.: Knowledge discovery and text mining
9. Goodwin, M.: The Roots of Extremism: The English Defence League and the Counter-Jihad Callenge. Chatham House (2013)
10. Hersovici, M., Jacovi, M., Maarek, Y.S., Pelleg, D., Shtalhaim, M., Ur, S.: The shark-search algorithm. an application: tailored web site mapping. Computer Networks and ISDN Systems 30(1), 317–326 (1998)
11. McNamee, L.G., Peterson, B.L., Peña, J.: A call to educate, participate, invoke and indict: Understanding the communication of online hate groups. Communication Monographs 77(2), 257–280 (2010)
12. Peng, F., Schuurmans, D., Wang, S.: Language and task independent text categorization with simple language models. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, vol. 1, pp. 110–117. Association for Computational Linguistics (2003)
13. Rawat, S., Patil, D.R.: Efficient focused crawling based on best first search. In: 2013 IEEE 3rd International Advance Computing Conference (IACC), pp. 908–911 (February 2013)
14. Reid, E., Chen, H.: Internet-savvy us and middle eastern extremist groups. Mobilization: An International Quarterly 12(2), 177–192 (2007)
15. Salem, A., Reid, E., Chen, H.: Content analysis of jihadi extremist groups' videos. In: Mehrotra, S., Zeng, D.D., Chen, H., Thuraisingham, B., Wang, F.-Y. (eds.) ISI 2006. LNCS, vol. 3975, pp. 615–620. Springer, Heidelberg (2006)
16. Sureka, A.: Mining user comment activity for detecting forum spammers in youtube. arXiv preprint arXiv:1103.5044 (2011)
17. Sureka, A., Kumaraguru, P., Goyal, A., Chhabra, S.: Mining youTube to discover extremist videos, users and hidden communities. In: Cheng, P.-J., Kan, M.-Y., Lam, W., Nakov, P. (eds.) AIRS 2010. LNCS, vol. 6458, pp. 13–24. Springer, Heidelberg (2010)
18. Ting, I.-H., Chi, H.-M., Wu, J.-S., Wang, S.-L.: An approach for hate groups detection in facebook. In: Uden, L., Wang, L.S.L., Hong, T.-P., Yang, H.-C., Ting, I.-H. (eds.) The 3rd International Workshop on Intelligent Data Analysis and Management. Springer Proceedings in Complexity, pp. 101–106. Springer, Netherlands (2013), http://dx.doi.org/10.1007/978-94-007-7293-9_11
19. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., Edwards, L.: Detection of harassment on web 2.0. In: Proceedings of the Content Analysis in the WEB, vol. 2 (2009)
20. Zhou, Y., Reid, E., Qin, J., Chen, H., Lai, G.: Us domestic extremist groups on the web: link and content analysis. IEEE Intelligent Systems 20(5), 44–51 (2005)