ABSTRACT

Online video sharing platforms such as YouTube contain 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.

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, over 6 billion hours of video are watched each month on YouTube. 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. 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, video spam, pornographic content, hate promoting and copyright infringed videos.

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. Previous studies show that extremist groups put forth hateful speech, offensive comments and messages focusing their mission. 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. 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

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Research Study</th>
<th>Platform</th>
<th>Objective &amp; Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>O’Callaghan et al; 2013</td>
<td>MB</td>
<td>Analysis of extreme right activities on multiple platforms for community detection.</td>
</tr>
<tr>
<td>2.</td>
<td>I-Hsien Ting et al; 2013</td>
<td>SN</td>
<td>Identifying extremist groups on Facebook using keywords and social network structure.</td>
</tr>
<tr>
<td>5.</td>
<td>P. Wadiwa et al; 2013</td>
<td>MB, VS, SN</td>
<td>Dynamic tracking of radical groups on web 2.0 by analyzing messages and posts.</td>
</tr>
<tr>
<td>6.</td>
<td>H. Chen et al; 2012</td>
<td>DF, VS</td>
<td>Examine several dark web forums and videos used by terrorist &amp; extremist groups.</td>
</tr>
<tr>
<td>7.</td>
<td>D. Deenning et al; 2012</td>
<td>VS, SN</td>
<td>An in-depth research on Social Media associated with jihad and counter terrorism.</td>
</tr>
<tr>
<td>9.</td>
<td>S. Mahmood; 2012</td>
<td>MB, VS, SN</td>
<td>Comparing several defence mechanisms to detect terrorists on social network websites.</td>
</tr>
<tr>
<td>11.</td>
<td>O’Callaghan et al; 2012</td>
<td>MB, VS, SN, OW</td>
<td>Activity and links analysis of extremist right groups (local and international) on Twitter.</td>
</tr>
<tr>
<td>12.</td>
<td>E. Kerez; 2011</td>
<td>DF</td>
<td>Quantitative and qualitative assessments of the content of communications on forums.</td>
</tr>
<tr>
<td>13.</td>
<td>D. David et al; 2011</td>
<td>MB, SN</td>
<td>Detecting criminal groups &amp; most visible players using the keyword search &amp; contacts.</td>
</tr>
<tr>
<td>14.</td>
<td>A. Sureka et al; 2010</td>
<td>VS</td>
<td>Locating hate promoting videos, users and their groups sharing a common agenda.</td>
</tr>
<tr>
<td>15.</td>
<td>L. McNamme et al; 2010</td>
<td>OW</td>
<td>Examining and comparing hate group websites to find their propaganda in messages.</td>
</tr>
<tr>
<td>17.</td>
<td>T. Fu et al; 2009</td>
<td>VS</td>
<td>Classifying extremist videos in on-line video portals using user generated metadata.</td>
</tr>
<tr>
<td>18.</td>
<td>Berrington et al; 2009</td>
<td>VS</td>
<td>Analyzing and distinguishing male &amp; female members in jihad radicalization groups.</td>
</tr>
<tr>
<td>19.</td>
<td>G. L. Muller et al; 2008</td>
<td>SN</td>
<td>Extracting key members and communities in social media for given dark web topics.</td>
</tr>
<tr>
<td>23.</td>
<td>E. Reid et al; 2005</td>
<td>DF, VS, OW</td>
<td>Mining U.S domestic and middle eastern extremist groups using hyperlinks.</td>
</tr>
<tr>
<td>25.</td>
<td>A. Salem et al; 2006</td>
<td>VS, OW</td>
<td>Identifying and analyzing several videos, extremist groups and their attributes.</td>
</tr>
<tr>
<td>27.</td>
<td>Y. Zhou et al; 2006</td>
<td>OW</td>
<td>Monitoring and locating extremist groups sites and their hidden communities on web.</td>
</tr>
</tbody>
</table>

Table 1: Summary of Literature Survey of 28 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.
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.20

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 43.

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 10, 11, 12, 17.

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 36.

6. A. Salem et al. propose a multimedia and content based analysis approach to detect jihad extremist videos and the characteristics to identify the message given in the video 39.

In context to existing work, the study presented in this paper makes the following unique contributions:

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 as the size of n-gram for language modelling 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 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 & 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. 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 modelling 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 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 31. 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. 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 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. 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

4. https://developers.google.com/youtube/getting_started
**Input** 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 3 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:

### 3.1.1 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 its children to be traversed next [35]. 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

---

**Table 2: Categorization of Sample Terms Occurring in Exemplary Documents for Focused Crawler.**

<table>
<thead>
<tr>
<th>Category</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Important Dates</td>
<td>13th January, 26th January, 23rd March, 5th August, 14th August, 15th August, 21st September, 9th November, 3rd December, 25th December</td>
</tr>
<tr>
<td>Region</td>
<td>Hindustan, Pakistan, India, Kashmir, Hindustan, Lahore, Afghanistan, America, Turkey, Mumbai, Kashmir, Indonesia, Punjab</td>
</tr>
<tr>
<td>Religion</td>
<td>Islam, Muslim, Hindu, Allah, Sikh, Kashmir, Bangladeshi, America, Christian, Buddhist, Jewish, Aristarche, Christian, Indigenous Muslim</td>
</tr>
<tr>
<td>Political Terms</td>
<td>Conspiracy, Leadership, Democracy, Terror, Awami, Strike, Khalistan, Against, Rights, Partition, Corruption, Media, Revolution, Objectives, Rule, Party, League, Politics, Politician, Slogan, Coalition, Public, President, Secularism, Domestic, Congress, Election, Witnessed, Tribal, Radicals, persecuted, Youth</td>
</tr>
</tbody>
</table>
explore its 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 5 and build a training data set. Algorithm SSA is a recursive function which takes \( U \) as a seed input. Steps 4 and 5 extract all features for seed user \( U \) and compute its similarity score with training profiles using character n-gram and language modelling (using LingPipe API). Steps 6 to 8 represent the classification procedure and labelling of users as relevant or irrelevant depending upon the threshold measures.

Algorithm 1: Focused Crawler- Best First Search

Data: Seed User \( U \), Walk of Graph \( w \), Size of Graph \( s \), Threshold \( th \), N-gram \( Ng \), Positive Class Channels \( U_p \)
Result: A connected directed cyclic graph, Nodes:User \( u \)

for all \( u \in U_p \) do
    \( D.add(\text{ExtractFeatures}(u)) \)
end

while \( 
\text{SizeOfGraph}(U) < s \) do
    \( 
    \text{u.getFriends()} \leftarrow \text{ExtractFeatures}(U) \\
    \text{score} \leftarrow \text{LanguageModeling}(D, U_f, Ng) \\
    \text{if} (\text{score} > th) \text{then} \\
        \text{U.class} \leftarrow \text{Relevant} \\
        \text{Usorted.InsertionSort}(U, \text{score}) \\
    \text{else} \\
        \text{Usorted.InsertionSort}(U, \text{score}) \\
    \text{end} \\
    \text{for } i = 1 \text{ to } u \text{ do} \\
        \text{Usorted.InsertionSort}(U) \\
        \text{end} \\
    \text{end} \\
    \text{for all } U_f \in U_f \text{ do} \\
        \text{fr} \leftarrow \text{ExtractFrontiers}(U_f) \\
        \text{Usorted.InsertionSort}(fr) \\
        \text{end} \\
    \text{end} \\
end

Algorithm 2: Focused Crawler- Shark Search

Data: Seed User \( U \), Walk of Graph \( w \), Size of Graph \( s \), Threshold \( th \), N-gram \( Ng \), Positive Class Channels \( U_p \), Decay Factor \( d \)
Result: A connected directed cyclic graph, Nodes:User \( u \)

for all \( u \in U_p \) do
    \( D.add(\text{ExtractFeatures}(u)) \)
end

while \( \text{SizeOfGraph}(U) < s \) do
    \( 
    \text{u.getFriends()} \leftarrow \text{ExtractFeatures}(U) \\
    \text{score} \leftarrow \text{LanguageModeling}(D, U_f, Ng) \\
    \text{if} (U \text{ is a child of irrelevant node}) \text{then} \\
        \text{score} \leftarrow \text{score} \cdot d \\
    \text{end} \\
    \text{if} (U \text{ has appeared before}) \text{then} \\
        \text{score} \leftarrow \text{max} \text{(new score, old score)} \\
    \text{else} \\
        \text{Usorted.InsertionSort}(U, \text{score}) \\
    \text{end} \\
    \text{for all } U_f \in U_f \text{ do} \\
        \text{fr} \leftarrow \text{ExtractFrontiers}(U_f) \\
        \text{Usorted.InsertionSort}(fr) \\
        \text{end} \\
    \text{end} \\
end

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 3. 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.

Algorithm 3: Features Extraction for a YouTube User

Data: User \( u \)
Result: User Activity Feeds and Profile Information

Algorithm ExtractFeatures(\( U \))
1. \( \text{uprofile} \leftarrow \text{u.getProfile()} \)
2. \( \text{uploads} \leftarrow \text{u.getUploadedVideos()} \)
3. \( \text{Commented} \leftarrow \text{u.getCommentedVideos()} \)
4. \( \text{Shared} \leftarrow \text{u.getSharedVideos()} \)
5. \( \text{favorited} \leftarrow \text{u.getFavVideos()} \)

3.1.2 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. [21]. 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 \( \text{score} \cdot \text{child} \cdot d \), 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 its similarity score by the maximum value of old and new inherited score. Steps 10 to 12 represent the classification procedure and labelling 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 3 and repeats steps 3 to 19 for each linked user.

3.1.3 Features Extraction

In Algorithm 3, we extract 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.

3.1.4 Frontiers Extraction

In Algorithm 3, 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
times for both BFS and SSA. Table 5 (a) and (b) present the seeds and each seed for all 6 modes. (BFS and SSA), we run our focused crawler 60 times for 10 values and 2 different n-gram values for similarity computation. 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.

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

4.1.1 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 4 shows a list of 35 user ids used as a training profiles. The 35 user ids consists of 612 videos and hence the training is performed on 612 videos. We obtain the training dataset by manually searching (keyword 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, favourited, 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.

4.1.2 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 3 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.

4.2 Experimental Results

4.2.1 Focused Crawler Results

As mentioned above, we execute our focused crawler 60 times for both BFS and SSA. 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, 1st quartile and 3rd 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 users (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 8 we see that the first quartile for mode A is below the threshold value and it is smaller than third quartile unlike in Figure 3. 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.
Figure 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

Figure 3: Box-Plot and Descriptive Statistics for Six Different Configurations of Best First Search Crawler

Figure 4: Box-Plot and Descriptive Statistics for Six Different Configurations of Shark Search Crawler
Table 5: Results of Focused Crawler for 10 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

Table 6: Confusion Matrix for Focused Crawler

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th>Relevant</th>
<th>Irrelevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant</td>
<td>991</td>
<td>295</td>
<td></td>
</tr>
<tr>
<td>Irrelevant</td>
<td>55</td>
<td>29</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>TPR</th>
<th>TNR</th>
<th>PPV</th>
<th>NPV</th>
<th>F1-Score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFS</td>
<td>0.75</td>
<td>0.35</td>
<td>0.88</td>
<td>0.18</td>
<td>0.81</td>
<td>0.69</td>
</tr>
<tr>
<td>SSA</td>
<td>0.77</td>
<td>0.35</td>
<td>0.95</td>
<td>0.09</td>
<td>0.85</td>
<td>0.74</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Network</th>
<th>NN</th>
<th>NE</th>
<th>SL</th>
<th>Dia</th>
<th>AD</th>
<th>ACC</th>
<th>IBC</th>
<th>ICC</th>
<th>#WCC</th>
<th>#SCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFS</td>
<td>23</td>
<td>119</td>
<td>3</td>
<td>4</td>
<td>0.225</td>
<td>0.388</td>
<td>0.046</td>
<td>0.356</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>SSA</td>
<td>24</td>
<td>137</td>
<td>8</td>
<td>3</td>
<td>0.238</td>
<td>0.788</td>
<td>0.069</td>
<td>0.320</td>
<td>1</td>
<td>16</td>
</tr>
</tbody>
</table>

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 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 324 (295 + 29) as irrelevant users. This misclassification of 22.92% and 65.47% 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 (921 + 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 fo-
4.2.2 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 represents a relation (friend, subscriber, and featured channel) between two users. Table 8 illustrates 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, therefore we are able to locate more number of hate-promoting communities in SSA. The average clustering coefficient of SSA is reasonably higher than BFS. Hence the clusters formed in SSA include only highly relevant users.

Figures 5(a) and 5(b) (generated using ORA+) 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. Colours of nodes represent the different in-degree of users and the width of an edge is scaled based upon the number of links between two users. In Figures 5(a) and 5(b) node A is the seed user. 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 in Figures 5(a) and 5(b) are different representation of the output graph based upon the betweenness centrality. 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. 

http://www.casos.cs.cmu.edu/projects/ora/
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 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 centre of all communities and connected to all users.

4.2.3 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 10. We categorize these videos based upon three main parameters: 1) focus of the content shown in the video, 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 10 reveals that the average duration of these videos is from 3 minutes to 45 minutes. Keywords present getting some specific audiences. Majority of videos are very large in the duration (3 to 45 minutes). Keywords present in the video. Table 9 shows that now users have used an-imates in front of their audiences. We divide these 274 videos into 12 categories based upon the type of content shown in the video. Table 9 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.

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.
6. REFERENCES


