Classification of YouTube Metadata using Shark Algorithm

Shubhangi D. Raverkar
MEPT (CSE) student
Department of CSE,
Government College of Engineering,
Aurangabad.

Meghana Nagori
Assistant Professor
Department of CSE,
Government College of Engineering,
Aurangabad.

ABSTRACT

YouTube is one of online video sharing platform that contains several videos and users promoting hate and extremism .Because of low barrier to publication and anonymity, YouTube is misused as a platform by most of users and communities to post negative videos spreading hatred against a particular religion, country or person. The problem of finding out of such hatred videos is proposed in this paper. 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 classier 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 comparer, 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.[1].

General Terms

Algorithm, shark search..

Keywords

YouTube metadata, Social Network Analysis, Hate and Extremism Detection, online radicalization

1. INTRODUCTION

Sharing video website mostly YouTube is used by users to upload an unlimited number of videos and access them, It allows to interact with each other by performing many social networking activities. As per YouTube statistics1 around 6 billion hours of video are watched each month. Billions 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 users2. The high reach ability of videos for users low publication barriers has led users to misuse YouTube in many ways by uploading malignant content that are offensive & illegal. Videos like harassment, insulting, video spam , pornographic content , hate promoting and copyright infringed videos . YouTube has become a convenient platform for lots of hate and extremist groups to share information and promote ideologies.

Video is the most usable medium to share views with other. Previous studies show that extremist groups put hateful speech, offensive comments, and messages focusing their mission. Social networking allows users (uploading extremist videos, posting violent comments, subscribers f these channels) to facilitate recruitment, gradually reaching worldwide viewers, connecting to other hate promoting groups, disseminating extremist content and forming their communities sharing a common agenda. Online radicalization and extremism have a major impact on society that contributes to the crime against humanity. 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 top such promotion in country). However, despite several community guidelines and administrative efforts made by you Tube, it has become a repository of lots of malicious and offensive videos. Detecting such hate promoting videos and users is significant and it is technically challenging problem. 100 hours of videos are uploaded every minute, that makes YouTube a very dynamic website. Hence, locating such users is overwhelmingly impractical..The aim is 1)to find out such videos and users, promoting hate and extremism (Focus of this paper) on YouTube, 2) to locate virtual and hidden communities of hate promoting users sharing a common goal or group mission 3) to find users with strong connections and playing central(or vital)l role in a community.

2. LITERATURE SURVEY

Some of the existing systems are summarized below which played vital role in finding out malicious videos, offensive languages, messages by using their methods like classification , algorithms, Learning approach etc.

In this closely related work to this paper is studied. Literature survey is given as per shown in Table 1 .Table shows study covered in pornography , cyber bulling, hate promoting videos ,offensive language detection & so on .

Table 1. Literature Survey of 8 Papers

Sr No	Author Name	Method	Aim
1.	Christopher C. Yang , Tobun D. Ng(2007)	a framework to analyze and visualize Weblog social networkis given. Link analysis uses the relationships between bloggers to construct the Weblog social network.[9]	Link, Content Analysis and Information Visualization
2.	Dawai Yin , Zhenzhen Xue , Liangjie Liony (2009)	Supervised Learning[10]	Detecting Harassment on Web 2.0
3.	Ying Chen, Sencun Zhu,Yilu Zhou,Heng Xu(2012)	Lexical Syntactic feature architecture[5]	Detecting Offensive Language in Social Media to Protect Adolescent Online Safety
4.	Vidushi Chaudhary , Ashish Surekha(2013)	Contextual Feature based one class classifier[8]	Contextual Feature Based One-Class Classifier Approach for Detecting Video Response Spam on YouTube
5.	Nilesh J.Uke, Dr. Ravindra C. Thool(2013)	Proposed system consists of three phases. Segmentation. Amount of nudity will be detected from rapid moving object detection phase and classification phases.[4]	Detecting Pornography on Web to Prevent Child Abuse – A Computer Vision Approach
6.	Vinita Nahar, Xue Li, Chaoyi Pang(2013)	approach to detect cyberbullying messages from social media through a weighting scheme of feature selection .A graph model is used to extract the cyberbullying network,.[3]	Cyberbullying Detection
7.	Swati Agrawal ,Nisha Agarwal, Ashish Surekha (2014)	One class classifier Approach [2]	Mining YouTube Metadata for Detecting Privacy Invading Harassment and Misdemeanor Videos
8.	Swati Agarwal, Ashish Sureka(2014)	best first search & shark search[1]	A Focused Crawler for Mining Hate and Extremism moting Users, Videos and mmunities on uTube”,2014 on “Best – t search and shark search”

3. RELATED WORK

In this closely related work to this paper is studied . Literature survey is given as per shown in Table 1 .Table shows study covered in pornography , cyber bulling, hate promoting videos ,offensive language detection & so on

1. Nisha Aggarwal et. al purpose an architecture for mining hate and extremisim promoting users,videos and Communities on YouTube.[1]
2. Nisha Aggarwal et. al. studied mining of YouTube metadata for detection of privacy invading harassment and misdemeanor videos. using one class classifier approach[2].
3. Vinita Nahar et. al. gave approach to detect t cyberbullying messages from social media through a weighting scheme of feature selection .A graph model is used to extract the cyberbullying network, which is used to identify the most active cyberbullying and through ranking algorithms.[3]
4. Nilesh J.Uke et. al. given system for Detecting Pornography on Web to Prevent Child Abuse byYing Chen et.al. given architecture Detecting Offensive Language in Social Media for online safety using lexical syntactic feature architecture.[5]As per existing work, the study

presented in this paper makes the following contributions:

1. this paper presents the study on adaptation of focused crawler framework (shark search) for navigating nodes and links on YouTube.
2. A series of experiments will be conducted on real-world data downloaded from YouTube to demonstrate the effectiveness of the proposed solution approach by varying some algorithmic parameters .
3. Social Network Analysis (SNA) based techniques will be applied on the retrieved user profiles and their connections obtained to understand presence of communities and central users.

4. PROPOSED SYSTEM

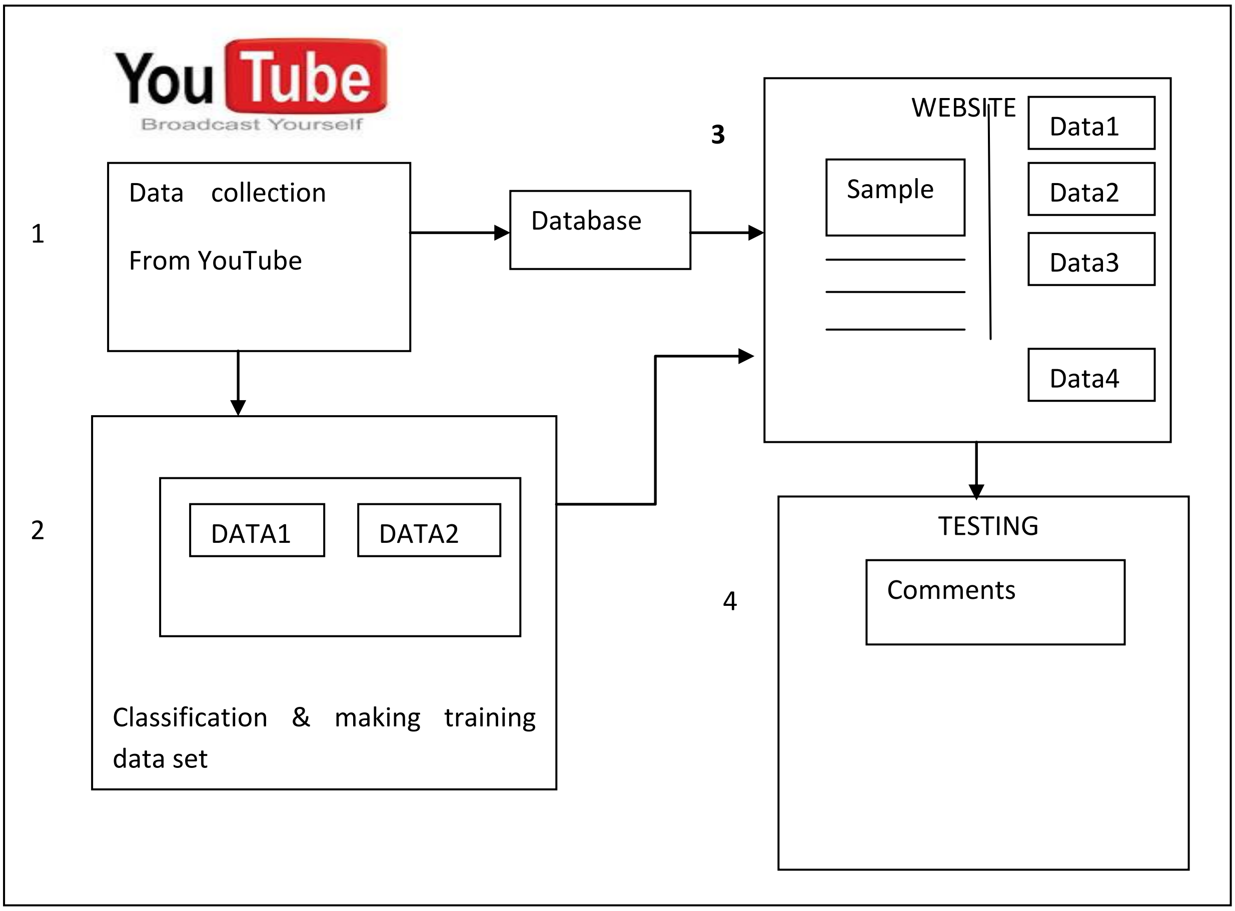

Figure 1 . A General architecture of proposed system

Figure 1 presents general solution to our proposed system. A proposed method is having four stages .Collection of Data, Training Data set, Web portal, Testing of Data given as stages 1,2,3,4 respectively.

The videos from multiple channels of YouTube like music, sports, Gaming, movies, TV shows, News, Live, Spotlight etc will be downloaded . This is stage 1. In stage 2

the manual analysis on collected data will be performed and make training set. There will collection of around 1000 videos (promoting hate and extremism) that will be used as training m and classified them accordingly. In stage 3, the web portal will be created. It takes one YouTube channel as a seed (a positive class channel) and extract it's metadata (user activity feeds and profile information) using YouTube API [6]. In stage 4 the extent of textual similarity between these metadata and training data is found by classification algorithm –shark search.

5. RESEARCH METHOD

Shark Search

Data: Seed User SU, Width of Graph wg, Size of Graph sg, Threshold t, N-gram Ng, Positive

Class Channels Up, Decay Factor d

```
Result: A connected directed cyclic graph, Nodes=User u
for all u 2 Up do
    D:add(ExtractFeatures(u))
end
Algorithm SSA(SU)
while graphsize < sg do
    userfeeds Uf  ExtractFeatures(SU)
    score score  LanguageModeling(D, Uf , Ng)
    if (SU is a child of Irrelevant node) then
        score   score _ df
    end
    if (SU has appeared before) then
        score   max(new score; old score)
    end
    if (score <t) then
        SU:newclass  Irrelevant
    else
        SU:newclass  Relevant
    end
    Hashmap Usorted:InsertionSort(SU; score)
    for i   1 to wg do
        Hashmap SUgraph:add(SUsorted(i))
    end
    for all SUg 2  SUgraph do
        fr = Extract Frontiers(SUg)
        Hashmap Ucrawler:add(fr)
    end
    for all Ufr 2 Ucrawler do
        SSA(Ufr)
    end
end
```

In the above algorithm comments of users will be as a input to function. The features will extracted from comments, language modeling will be done by using threshold value t. The data is classified.

6. CONCLUSION

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 le compared with trained data which is done by manual analysis.

7. REFERENCES

[1] Swati Agarwal, 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"

[2] Nisha Aggarwal, Swati Agrawal, Ashish Sureka "MiningYouTube Metadata for Detecting Privacy InvadingHarassmentandMisdemeanorVideos",2014.on"oneclass classifier".

[3] Vinita Nahar , Xue Li, Chaoyi Pang "An Effective Approach for Cyberbullying Detection" 2013.

[4] Nilesh J.Uke, Dr. Ravindra C. Thool "Detecting Pornography on Web to Prevent Child Abuse – AComputer Vision Approach "2013

[5] Ying Chen, Sencun Zhu,Yilu Zhou,Heng Xu "Detecting Offensive Language in Social Media to ProtectAdolescent Online Safety "2013

[6] A. Bermingham, M. Conway, L. McInerney, N. O'Hare, and A. Smeaton. Combining socialnetwork analysis andsentiment analysis to explore thepotential for onlineradicalisation. In Social Network Analysis and Mining, 2009. ASONAM '09.International Conference on Advances in, pages231{236, 2009.)

[7] April Kontostathis,Kelly Reynolds,Andy Garron,Lynne Edwards "Detecting Cyberbullying: Query Terms and Techniques"2013.

[8] Vidushi Chaudhary , Ashish Surekha"Contextual FeatureBased One-Class Classifier Approach for Detecting Video Response Spam onYouTube" (2013)

[9] Christopher C. Yang , Tobun D. Ng"Terrorism and CrimeRelated Weblog Social Network:Link, Content Analysis and Information Visualization(2007)

[10] Dawai Yin , Zhenzhen Xue , Liangjie Liony" Detectionof Harassment on Web 2.0" 2009.