Semi-Automated System For Filtering Objectionable Content Using Video Metadata

Vijay Kumar S, Manjunath G S

Abstract: The most popular and widely accepted video sharing websites on Internet are YouTube, Vimeo, Metacafe, Veoh, The Internet Archive, Crackle, etc. Ample numbers of videos on these platforms have unwanted content, so mining of video metadata can be employed to identify such videos. Upon downloading several videos as dataset and manually explaining the dataset will help in studying and categorizing training dataset. The detection of obnoxious videos and frame in the dataset can be classified using "One class classifier approach". Image thumbnails are used instead of video thumbnails for efficient bandwidth consumption. Offline mode is enhanced by providing full-download option to the users, without having to re-synchronization periodically. Integration with YouTube is a feature included to play the selected video content.

Index Terms: video sharing websites, objectionable video filtering, metadata, stemming, Ranking, youtube integration APIs

1 INTRODUCTION

The Video sharing plays an essential role in the field of education in recent times. Sites such as YouTube, Vimeo, Metacafe, Veoh, The Internet Archive, Crackle, etc. are few of the most popular and widely accepted video sharing websites on Internet. Video sharing websites provide services such as file hosting, image hosting and social networking services. Due to the low publication barrier and anonymity in youtube many objectionable contents are uploaded in significant percentage by violating YouTube community guidelines. It also contains numerous dishonored videos, spam, negative and activism promoting videos, objectionable material and privacy assaulting contents are also uploaded. This work presents an approach to identify privacy invading harassment and misdemeanor videos by mining the video metadata. A one class classifier approach is used to detect the objectionable video and frames. The analysis of test dataset reveals that semantic features can be used to predict the video type is done. Upon conduction of series of experiments on evaluation dataset acquired from youtube the validation of hypothesis is done for the proposed approach to check its accuracy. For a given keyword the search engine returns thousands of results extracted from the surrounding text from many web sources. When users search for a query on any social media website, numbers of videos are displayed. Once it is moved to other pages provided by filtering approaches, number of unrelated videos will be present that occupies large bandwidth, this causes spam to the users and leads to deviation from intended information. Primarily the video presenting or hosting services are provided by websites or software which for meant for this purpose only. This hosting service allows users to distribute their video clips. The primary video sharing services are supported by other websites and services such as file hosting services, image hosting services and social network services. Many websites provide services like sharing data privately and publish them. The video hosting services can be classified into several categories like: video sharing websites/ platforms, white label providers and web-based video editing sites. Few websites are solely works as a search engines and they do not disclosure their video content (such as singing fish) are not considered for this work. Some services chargeable, but most are available for free. Some websites offer pay-per-view for their videos as commercialization features. Most of the user generated sites offer free services whereby users can upload video clips and allow others to access in a large group. Many sites have a Terms of Service information and can take judgment calls on contents uploaded by qualifies, along with that they place restrictions on the file size, duration, content to be uploaded and format of the uploaded video file. Some sites verify user age before providing access control to adult material. Some sites apply primary filters and clean the content before it is published, then community of users considered as a "reviewer" filter out inappropriate contents. There are different kind of users categorized based on the content they share knowingly or unknowingly. In order cause discomfort for the victim few users knowingly post videos on YouTube that threatens and disturb one or more people. For example, violent, abusive and humiliating behaviour that violates the claimant's dignity. Sometimes users take a clip of some incident and share it on YouTube without any intention to hurt that person involved in the video. The above activity has an impact on the world in both positive and negative way. The effect of YouTube and all that comes with it forms negative violence, hurtful thing, and cruelty, the list keeps going on a high negative effect. The negative effects on animals that YouTube is now a part of has increased in numbers. People get amusement from suffering and pain such action is turning us into a heartless culture instead of helping the innocent.

2 RELATED WORK

Danushka Bollegala, et.al, [1] in "Automatic Discovery of Personal Name Aliases from the Web" had proposed that every individual is typically referred by many numbers of name aliases on the web. For sentimental analysis, information retrieval, personal name disambiguation and relation extraction tasks can be done using Aliases identification through a given person name is found accurate and useful in various web related tasks. A new method to extract aliases of a given personal name from the web is proposed. The proposed method accepts personal name then first extracts a set of candidate aliases. Second, rank the extracted candidates according to the likelihood of a candidate being a correct alias of the given name. An approach automatically extracts lexical pattern-based approach to efficiently extract a large set of candidate aliases from snippets retrieved from a web search engine is proposed. Reiner Kraft, et.al, [2] in "Mining anchor text for query refinement" had proposed that, Searching the large hypertext document collection is often possible that there are too many results available for ambiguous queries, to overcome the issue query refinement is required. An interactive process of query modification that can be used to narrow down the scope of search results is called " Query refinement".A new method for automatically generating

refinements or related terms to queries by mining anchor text for a large hypertext document collection is proposed. The proposed approach suggests that text refinement can also be used to augment legacy query refinement algorithm based on logs of query, since they typically differ in coverage and produce different refinements. The results are based on experiments on an anchor text collection of a large corporate intranet. TaikiHonma et.al, [3] in "Identification of Actual Name on the Web" had proposed that, the designing of robust alias detection method needs ranking scores that are integrated with page-count-based association measures done using support vector machines. The proposed method by the author outperforms with respect to other numerous baselines and previous work carried out on alias extraction on a dataset of personal names. For achieving a statistically significant mean reciprocal rank many experiments are carried out using a dataset of location names and Japanese personal names. This experiment helped in suggesting that possibility of ex-tending the proposed method to extract aliases for different types of named entities and for other languages. The proposed method improves recall by 20% in a relation-detection task using aliases extraction. G. Tireesha Kumari et.al, [4] in "Detection of Name Disambiguation and Extracting Aliases for the Personal Name" had proposed that an individual can be referred by multiple name aliases on the web. For efficient information retrieval, sentiment analysis and name disambiguation, extracting aliases of a name is most important. Authors have proposed a novel approach to find aliases of a given name using automatically extracted lexical pattern-based approach.

The following approach is used to for defining an efficient ranking score to evaluate candidate aliases:
The word co-occurrences in an anchor text and page counts on the web.
The mutual relations between words that appear in anchor text, words in anchor text are represented as nodes in the co-occurrence graph and edge is formed between nodes which link to the same URL.

Sureka, A.Kuma raghu, et.al, [5] "Mining youtube to discover extremist videos, users and hidden communities in information retrieval technology" had proposed that the focus of this work is on data mining in YouTube to discover the amount of hate videos, users and virtual hidden communities. Since YouTube repository is increasing day to day finding precise information on YouTube is a challenging task.

J. Golbeck, et.al, [6] in "Combining provenance with trust in social networks for semantic web content filtering" had proposed that in Data mining the Classification is a data mining function that allocates similar data to categories or classes. One of the most common methods for classification is ensemble method which refers a machine learning technique supervised learning. After generating classification rules one can apply those rules on unknown data and reach to the results. In one-class classification it is assumed that only information among one of the classes, the target class, is available. This means that just example objects of the target class can only be used and that no information about the other class of outlier objects is exist.

In One Class Classification (OCC) problem solving technique the negative class is either absent or improperly sampled. There are different classification mechanisms that can be used. In an ensemble classification system, different base classifiers are combined in order to obtain a classifier with better performance. The popularly used ensemble learning algorithms are AdaBoost and Bagging. The method of ensemble learning method can be divided into three phases: the generation phase, in which a set of candidate models is induced, the pruning phase, to select of a subset of those models and the integration phase, in which the output of the models is combined to generate a prediction. Keywords: Bagging, Boosting, Classification, Ensembles, One Class Classification, Positive and Unlabeled Data.

H. Sch·u·tze, et.al, [7] "A comparison of classifiers and document representations for the routing problem" had proposed that in today's day, the Internet provides infinite amount of data. A lot of this data is useless. From the analysis perspective the useful data can be mined and used. The data provided from the twitter forum is used to predict the occurrence of any mishap. Sentiment analysis is used to find out the peoples' reaction to certain events. Based on the reaction of many people towards the objectional content with different views are present in close locations there is a possibility of a crime. The web provides volumes of text-based data which are stored in online chatting websites like Twitter, Face book, Blog and Forum etc. Cyber bullying is a socially aggressive and has powerful negative effects for individuals, specifically adolescents and youngsters. In the recent times many methods for automatic thoughts of mining in the online data are becoming increasingly important, to increase the safety parameter of the people. This framework is proposed to extract Cyber bully polarity from the Forum using Fuzzy logic technique. First, the given input is pre-processed, and the useful content is gathered using data mining. Subsequently, the pre-processed data will be sent to the extract the features. A probability of the words is calculated by using Fuzzy Decision Tree Method. Fuzzy rules can be applied in all these features to extract the certain set of cyber bully words like bad words, insulting words, threatening words and terrorism words from the given input, hence text mining is used here. Finally, this method will return the reduced and accurate cyber bully words. This method is carried out by human annotation using the existing methods like Mamdani Fuzzy System and Naive Bayes classifier using AI. Extensive experiments are performed by using fuzzy logic on crime debate forum and the results shows that this proposed approach is better than the existing one. KEYWORDS: Cyberbully, text mining, Forum, fuzzy-logic, fuzzy decision tree, Naive Bayes classifier, tweets, feature extraction, R language, Geographical prediction. Twitter, Python, Sentiment analysis.

D. D. Lewis, et.al, [8] in "An evaluation of phrasal and clustered representations on a text categorization task" had stated that the action recognition problem was become a hot topic within computer vision, the detection of fights or in general aggressive behavior has been comparatively less studied. Video surveillance scenarios of prisons, psychiatric centers or embedded camera phones is useful to measure the capability. Recent work has considered for fight detection problem the well-known Bag-of-Words framework often used in generic action recognition. Using this framework, spatio-temporal features are extracted from the video sequences and used for classification of data. Despite encouraging results in

which near 90% accuracy rates were achieved for this specific task, the cost of computation in extracting such features is prohibitive in real-time for practical applications, particularly in surveillance and media rating systems. Specific features are leveraged to detect the violence. Psychological factors show that kinematic features alone are discriminated for specific actions, this work proposed a novel method which uses extreme acceleration patterns as the main feature. The extreme accelerations are efficiently estimated by applying the Radon transform to the power spectrum of consecutive frames. The level of accuracy improved and achieved up to 12% with respect to state-of-the-art generic action recognition methods. Most important thing is the proposed method is 15 times faster than existing.

Drawbacks of related systems

- Unnecessary display of unrelated videos
- Does not Eliminate the unwanted videos
- Bandwidth is wasted
- There is no control on the contents being posted
- Offline mode is not provided on the PC

Problem Statement

To develop a semi-automated system which filters videos containing objectionable content using metadata of videos, incorporates effective utilization of bandwidth and provides an offline functionality.

3 PROPOSED SYSTEM

A semi-automated system which can filter unwanted Videos from user walls is developed to make video sharing safer and more productive. The proposed system provides a powerful rule layer exploiting a flexible language to specify the Filtering Rules (FRs), by which users can state what contents, to be displayed and what should not be displayed. The proposed method will work on the sub-name and get the association orders between names and sub-name to help search engine tag that sub-name according to the orders such as first order associations, second order associations. The percentage of related documents are retrieved for a search option to query on search engine. The mean rank of the search engine for a sample set of queries is the average of reciprocal ranks for every query. The term co-occurrence refers to the temporal property of the two terms occurring at the same web page or same document on the web. The anchor text is the clickable text on web pages, which points to a web document. Tireesha Kumari, Mr. Saroj Kumar Gupta “Detection of Name Disambiguation and Extracting Sub-name for the Personal Name” An individual can be referred by multiple name sub-name on the web. Extracting sub-name of a name is important in information retrieval, sentiment analysis and name disambiguation. We propose a novel approach to find sub-name of a given name using automatically extracted lexical pattern-based approach.

Advantages of Proposed System

- The live integration is made possible with the help of YouTube API.
- Both user and admin can upload videos on to the local server.
- It detects normal and unwanted video. Then provides spontaneous filtering.
- Offline mode is a useful functionality provided on PC without a third-party application.
- It provides effective bandwidth usage with the help of image thumbnails.
- It provides security by user authentication.

Objectives of Proposed System

- Retrieve video of the important persons who has multiple sub-names.
- Administrator must be able to upload or delete the information.
- Architectural Re-Design is simple.
- Accuracy level must be high.
- Reliability of the system must be maximum.
- Performance of the system must be high.

4 SYSTEM ARCHITECTURE

The following figure shows system architecture of proposed system and how it will work on the aliases and get the association orders between name and aliases.

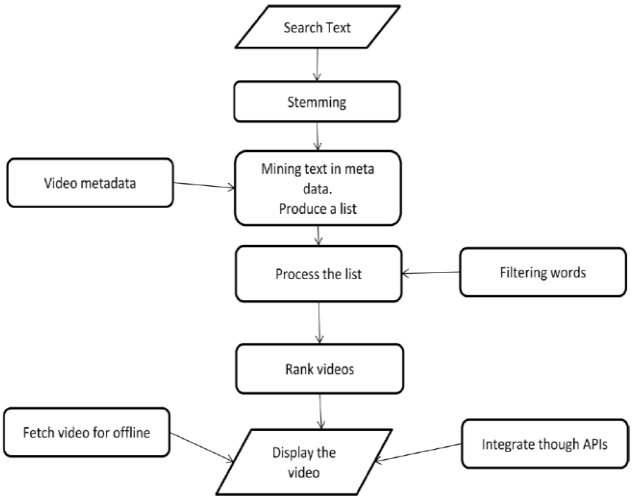

Fig. 1: *System Architecture*

The anchor texts which point to the same URL are called as inbound anchor texts.

- The anchor texts-based co-occurrences between name and aliases using co-occurrence statistics.
- Stemmer algorithm is a process for removing the commoner morphological and in flexional endings from words in English.
- Then a word co-occurrence graph will be created and mined by graph mining algorithm so as to get the hop distance between name and aliases that will lead to the association orders of aliases with the name.
- Elimination of unwanted videos
- Offline video
- A real time YouTube integration

5 MATERIALS AND METHODOLOGY

The proposed system implementation is done using JAVA and Web programming languages such as HTML, XML and CSS. The following soft wares are used to execute the project code i.e Netbeans, Navicat100 & JDK- 7u and database connective is done through Microsoft MySQL essentials. The proposed system uses the methodology which retrieves all the corresponding URLs for all anchor texts in which name and

aliases appear, then two websites are checked for the anchor texts co-occurrence and its frequency. The future vector is developed with the help of normalized values obtained from the training samples and algorithm specified in Implementation part of ranking module. The youtube integration is done for the videos using APIs and code written under youtube integration module of implementation. Finally the videos are filtered using stemming and morphology algorithms.

6 IMPLEMENTATION

Implementation phase generally consists of proper careful planning, investigation of the existing system and its constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods. Implementation is an important phase in the development of the project where the software design is realized as a set of program units. The objects that are identified in the design stage are implemented and functions, which manipulate these objects, are realized.

There are 6 modules included in the application and are listed as follows:

- Co-occurrences in Anchor Texts module
- Anchor Texts Co-occurrence Frequency module
- Ranking module
- Real time YouTube Integration module
- Filtration module
- Stemmer module

A Co-occurrences in Anchor Texts module

The proposed method first tries to retrieve all corresponding URLs from search engine for all anchor texts in which name and aliases appear. The existing search engines provide search operators to search in anchor texts on the web. For example, Google provides in anchor or Allin anchor search operator to retrieve URLs that are pointed by the anchor text given as a query. For example, query on "Allin anchor: Hidek Matsui " to the Google will provide all URLs pointed by Hideki Matsui anchor text on the web. The objective of the proposed search engine is to provide the most relevant documents for any user queries. Anchor texts play a vital role in search engine algorithm because it is clickable text which points to a relevant page on the web. The search engine considers anchor text as a main factor to retrieve all the relevant documents to the user's query. Anchor texts are also used in synonym extraction, ranking and classification of web pages and query translation in cross language information retrieval system.

B Anchor Texts Co-occurrence Frequency module

Two different web pages are generally used to check the two anchor texts appears in different page is called as inbound anchor texts co-occurrence. The co-occurrence frequency of anchor texts refers to the number of different URLs on which they occur.

C Ranking Module

The training samples to be normalized into the range of [0-1] for all the co-occurrence which measures the anchor texts. The normalized values termed as feature vectors will be used to train the module to get the ranking function to test the given anchor texts of name and aliases. Then for each anchor text, the trained module using the ranking function will rank the other anchor texts with respect to their co-occurrence measures with it. The highest-ranking anchor text will be elected from a lot to make a first-order association with its corresponding anchor text for which ranking is done. The co-occurrence graph is drawn for the name and aliases as per the first order associations between them.

Algorithm is as follows:

JDBC connection

String qry ="select* from new where id="'+ id+'"";
Session commenting
Updating count with the query while (rs.next())
{
count = count + 1;
}
Update final result
Display in descending order

D YouTube Integration module
Clicking on the videos makes integration with real time YouTube server and fetches the videos and displays it to the users.
1. YouTube Integration
2. Get session attribute (vname);
var tag = document.createElement('script'); tag.src = https:llwww.youtube .comliframe_ api";
varfirstScriptTag=document.getElementsByTagName('script')[0];
firstScriptTag.parentNode.insertBefore(tag, firstScriptTag);
3 This function creates an <iframe> (and YouTube player) after the API code downloads. var player;
Function on YouTube
lframeAPIReady()
{
var id=document.ff.val.value;
player= new YT.Player('player',
{
height: '390',width: '620', videold: id, events:
{
'onReady':onPlayerReady,'onStateChange':
onPlayerStateChange
}
}
}

4.The API will call this function when the video player is ready.
functiononPlayerReady(event) {
event.target.playVideo();
}

5. The API calls this function when the player's state changes. The function indicates that when playing a video (state=l), the player should play for six seconds and then stop. var done = false;
functiononPlayerStateChange(event) {
if (event.data== YT.PlayerState.PLAYING&& !done) {
setTimeout(stopVideo, 6000);
done= true;
}
}
Function stop Video() { player.stop Video();

```
}
</script>
<form name="ff '>
<input type="text" name="val" id="val" value="<o/o=vnameo/o>"/><input type="submit" name=" submit" onclick="onYouTubeIframeAPIReady()"/>
```

E Filtration Module

This module is determines what content will be available or be blocked. Such restrictions can be applied at various levels: a government can attempt to apply them nationwide (see Internet censorship), or they can, for example, be applied by an ISP to its clients, by an employer to its personnel, by a school to its students, by a library to its visitors, by a parent to a child's computer, or by an individual user to his or her own computer. The motive is often to prevent access to content which is considered objectionable.

Content filtering software can, however, also be used to block malware and other content that is or contains hostile, intrusive, or objectionable.

```
Initialize String tag= request.getParameter("search");
Initialize String id = null, sname = null, vname = null,iname = null,tagl, null,tag2 =null;
ArrayList <String > words=new ArrayList<String>();
```

JDBC Connection

```
String qry ="select* from vulger ";//getting all filter word
while(rsl.next())
 {
tag2 = rsl.getString("name");
words.add(tag2);
}
```

JDBC connection

```
String qry = "select* from new where tag like'%" +tag+"%' or iname like '%"+tag+"%' order by comment DESC";
while (rs.next())
{
// mapping with result set if(!words.contains( iname))
{
//filtering
```

Display Result Data Set

F Stemmer module

The process of reducing inflected (or sometimes derived) words to their word stem, base or root form generally a written word form using syntactical morphology and information retrieval is used. The stem need not be identical to the morphological root of the word; it is usually enough that related words map to the same stem, even if this stem is not in itself a valid root.

A process called conflation which many search engines treat words with the same stem as a synonym kind of query expansion. Algorithms for stemming have been studying from computer science since from 1960s.

The Porter stemming algorithm (or 'Porter stemmer') is one of the methods used for removing the commoner morphological and in flexional endings from words in English. The main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems. Stemming programs are generally referred as stemming algorithms or stemmers. The algorithm is as follows:

Rule format-

The rules are of the form: (condition) S1 -> S2, where S1 and S2 are suffixes

Notations used are:

1. m - The measure of the stem
2. *S - The stem ends with S
3. *v* - The stem contains a vowel
4. *d - The stem ends with a double consonant
5. *o - The stem ends in CVC (second C not W, X, or Y)

Step 1:

- SSES -> SS e.g. caresses -> caress • IES -> I e.g. ponies -> poni, ties -> ti
- SS -> SS e.g. caress -> caress
- S -> ϵ eq. cats -> cat

Step 2:

- (m>1) EED -> EE
- Condition verified: agreed -> agree
- Condition not verified: feed -> feed
- (*V*) ED -> ϵ
- Condition verified: plastered -> plaster
- Condition not verified: bled -> bled
- (*V*) ING -> ϵ
- Condition verified: motoring -> motor
- Condition not verified: sing -> sing

Step 2b:

- (m>1) EED -> EE
- Condition verified: agreed -> agree
- Condition not verified: feed -> feed
- (*V*) ED -> ϵ
- Condition verified: plastered -> plaster
- Condition not verified: bled -> bled
- (*V*) ING -> ϵ
- Condition verified: motoring -> motor
- Condition not verified: sing -> sing

Step 3:

- Y Elimination (*V*) Y -> I
- Condition verified: happy -> happi
- Condition not verified: sky -> sky

Step 4: Derivational Morphology I

- (m>0) ATIONAL -> ATE
- Relational -> relate
- (m>0) IZATION -> IZE
- generalization-> generalize
- (m>0) BILITI -> BLE
- sensibiliti -> sensible

Step 5: Derivational Morphology II

- (m>0) ICATE -> IC
- triplicate -> triplic
- (m>0) FUL -> ϵ
- hopeful -> hope
- (m>0) NESS -> ϵ
- goodness -> good

Step 6: Derivational Morphology III

- (m>0) ANCE -> ɛ
- allowance-> allow
- (m>0) ENT -> ɛ
- dependent-> depend
- (m>0) IVE -> ɛ
- effective -> effect

Step 7a:

- (m>1) E -> ɛ
- probate -> probat
- (m=1 & !*o) NESS -> ɛ
- goodness -> good

Step 7b:

- (m>1 & *d & *L) -> single letter
- Condition verified: controll -> control
- Condition not verified: roll -> roll

7 EVALUATION AND RESULTS

The code implemented was executed with the help of software listed in methodology. The following are the sequence of actions performed during execution of code implemented.

1. The new users are allowed to register. It contains 4 fields: user id, password, gender, phone number. The submit button saves the details on database. A customer can have only one account which will be taken care by checking for repeated phone number. The saved details are later used to validate the user login credentials
2. The user login page authenticates the registered user. It contains two fields: user id and password. Both the fields are compulsory. An alert is generated if the fields are not filled by the user. When the submit button is clicked and the credentials are checked for validity against the existing accounts, the user home page is opened.
3. The users are allowed to search for the videos by typing in anchor text in the search tab. the anchor text is stemmed and used for searching. Stemming processes the word and converts the word to its root form.
4. The admin webpage designed allows the uploading of videos along with its metadata to the web server. There are 6 fields: upload id, parent name, sub name, video id, thumbnail and description. All the fields are compulsory.
5. Each uploaded video has the option to like and comment which is available under the “read more” option.. This page also displays the number of likes and comments for a video. Users are given the option to like and comment on the video which will be displayed to the other users as well, after refresh.
6. Each video which is uploaded is assigned with a video id. Video id is assigned when the videos are uploaded though the upload option. Admin is given the authority to delete videos. This id is used by the admin to delete the video from the list.
7. Administrator can monitor the content of the site by uploading words which are used for training the system to filter the videos.
8. Filtration uses the list of words which are predefined and searches for them in the metadata. If the words are found, the video is not displayed.
9. The list of videos which are fetched based on the matches and then checked for objectionable content. Only if the data is not objectionable, it is displayed to the user.
10. Finally the ranking of videos is done according to the most number of likes and comments are done in order to consider popularity among the videos hence making searches more efficient.

Based on the execution done for the code implemented following results are tabulated.

Sl. No	Anchor text used for searching	No. of Videos actually listed	No. of videos displayed for user	No. of videos deleted based on objectionable words
1.	were the twins towers brought down by explosives demolitions	25	14	11
2.	Terror activities	38	05	33
3.	People murdered	45	10	35

Table 1. *Sample result*

Screenshots

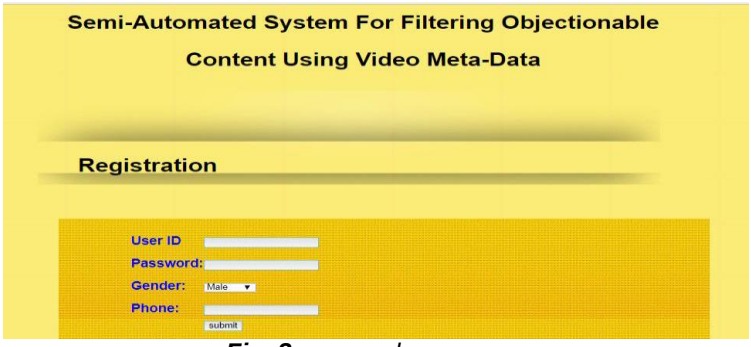

Fig. 2*. user web pages*

Fig.3*. Search Page after login*

Fig.4*. Page to Upload Filter Words for admin*

***Fig. 5**. The list of videos after filtering, as Viewed by user*

***Fig. 6**. The List of Videos as Viewed by Admin*

***Fig. 7**. The List of Videos as Viewed by user after ranking*

8 CONCLUSION

Developed a semi-automated system to identify privacy invading harassment and misdemeanour videos by mining the video metadata, included features like re-ranking based on number of comments and likes which is an add-on to video sharing applications. This application is objectionable-content-free, secured, easy to use and makes all its users happy. A portrayal study on a training dataset can be done by downloading several videos using YouTube API and manually annotating the dataset was conducted. Several discriminatory features for recognizing the target class objects are defined. A one class classifier approach to detect the objectionable video and frame the problem as a recognition problem was employed. In this system, ranking is based on number of likes and comments, ranking can also be done using other parameters such as views, description, author etc. Classification can be extended to include dynamic content from user's comments. These can be the future enhancements.

REFERENCES

[1] DanushkaBollegala, Yutaka Matsuo, and Mitsuru Ishizuka, "Automatic Discovery of Personal Name Aliases from the Web"
[2] Reiner Kraft,CAJasonZien"Mining anchor text for query refinement"
[3] TaikiHonma "Identification of Actual Name on the Web"
[4] G. TireeshaKumari, Mr. Saroj Kumar Gupta "Detection of Name Disambiguation and Extracting Aliases for the Personal Name"
[5] Sureka,A.Kumaraghu,P.Goyal,& Chhabra.S "Mining youtube to discover extremist videos,users and hidden communities in information retrieval technology" In the Year:2010
[6] J. Golbeck, "Combining provenance with trust in social networks for semantic web content filtering," in Provenance and Annotation of Data, ser. Lecture Notes in Computer Science, L. Moreau and I. Foster, Eds. Springer Berlin / Heidelberg, 2006, vol. 4145, pp.
[7] H. Sch¨utze, D. A. Hull, and J. O. Pedersen, "A comparison of classifiers and document representations for the routing problem," in Proceedings of the 18th Annual ACM/SIGIR Conference on Resea. Springer Verlag, 1995, pp. 229–237.
[8] D. D. Lewis, "An evaluation of phrasal and clustered representations on a text categorization task," in Proceedings of 15th ACM International Conference on Research and Development in Information Retrieval (SIGIR-92), N. J. Belkin, P. Ingwersen, and A. M.