iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** V **Month of publication:** May 2019

DOI: https://doi.org/10.22214/ijraset.2019.5067

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

Detection of Media Piracy using Artificial Intelligence, Machine Learning and Data Mining

Prof. Chaitanya Mankar[1], Kasturi Khedkar[2], Shubhangi More[3], Pavitra Phand[4], Poonam Pawar[5]
[1,2,3,4,5]*Department of Computer Engineering, Dhole Patil College of Engineering, Kharadi, Pune, India*

Abstract: ***We know that Web Technology is evolving continuously day by day which helps media creators in marketing and distribution of media content. Even though these technologies help the creator and the users, it will also have a high chance of increasing pirated contents distribution and redistribution. As people have so many different ways to share and distribute the web content like Free Cloud Spaces , Social Networking Portals, Email, Drives, Chats etc., so finding or detecting these contents manually is a difficult and time consuming task. We can make use of concepts like ML and AI to fight piracy by using content monitoring solutions. In this paper, it have been explained that how we can use the different techniques of DM, AI and ML to search the web contents and identify piracy threats which must be managed in and efficient way. Using different web services and the concept of machine learning, the system will produce data of the searched contents of media with various information like IP Addresses, source of piracy, time, region, period etc. Also the system is going to store or find out the blacklisted and untrusted websites and webpages which in turn can be used by different Private Organizations or Government Agencies to get rid of the Piracy.***

Keywords: Artificial Intelligence, Machine Learning, Data Mining, Information Extraction, Randomized Search, Supervised Learning, Data Cleaning.

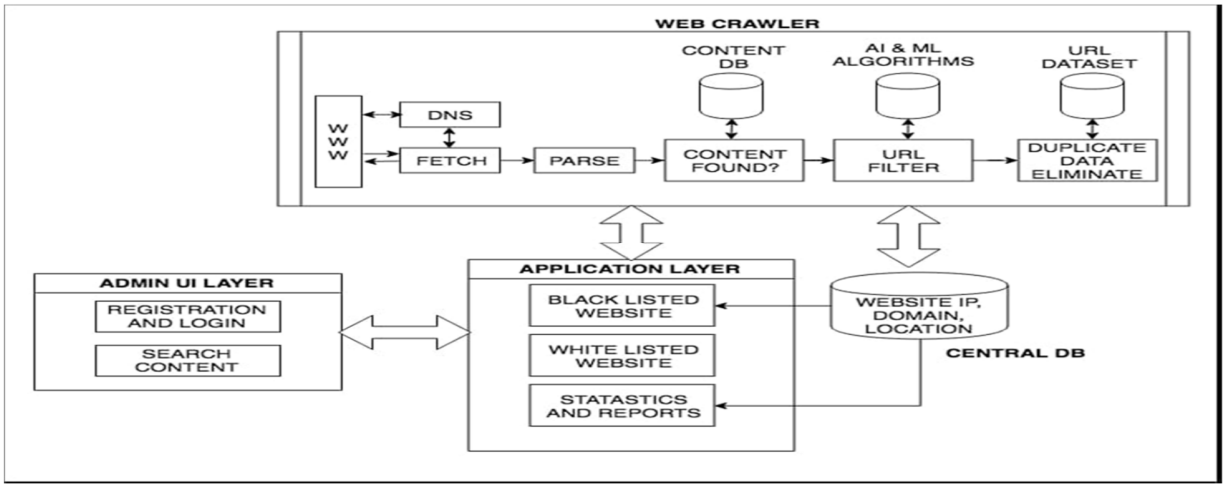
I. INTRODUCTION

Artificial intelligence is an intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans and animals. Machine learning is a field of Artificial Intelligence in which it studies algorithms and different statistical models which can be used by machines or computers to perform tasks without explicit instructions. Artificial Intelligence and Machine Learning can be used to fight piracy by using content monitoring solutions. Data mining is the process. It is used for discovering patterns in large data sets. Data mining techniques will be used to search the web for the media content and identify piracy threats. It will give data of searched media content. This data can be used by Private Organizations or Government Agencies to get rid of the Pirated content of the Web.

II. RELATED WORK/SYSTEM ARCHITECTURE

In the Working system or the Architecture, we have four modules Admin UI layer, Web Crawler, Central DB and Application layer. The Admin UI layer is responsible for the Admin and user Login. It also has the Data or the Database which is to be given to the Organizations or the Agencies. In the Web Crawler, it is used for searching web pages one by one by providing indexing, in this layer the DNS will be fetched and one by one parsed then contents will be searched in content DB. If the content is found in the content DB it is directly sorted as Blacklisted or Whitelisted and if the contains are not found in the Content DB then using AI and ML algorithm URL will be filtered, and after that it will be checked and then sorted as Blacklisted or Whitelisted. If the duplicate data is found then from the URL data set, duplicate data will be eliminated. In the Application layer list of Blacklisted and Whitelisted websites will be shown. In the central database, the websites IP Address with its location, domain will be stored and the website will be manually checked, from their websites will be categorized as Blacklisted and Whitelisted and accordingly statistics and reports will be generated and send to be Admin for further process.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com



A. Modules

1) *Admin Module:* It consists of mainly Reports. The report is generated from the updated database. It includes Black Listed Websites and White Listed Websites. The Admin Module is also responsible for managing users.
2) *Black Listed Websites:* Black Listed Websites consists of mainly those sites which are not legal, and from which we can download the content illegally those websites are displayed in the Black Listed Websites. This will be the Module of which the contents will be sent to the Government Agencies or Piracy Protection Organizations.
3) *White Listed Websites:* Those websites which are legal are the ones displayed under the White Listed Website. So, White Listed Websites consists of mainly those sites which are legal, and from which we can download the content legally.
4) *Database:* The Database also saves the newly searched data and helps in fast processing. All the White Listed and Black Listed websites are stored in the database.

III. PROPOSED ALGORITHM

A. Design Considerations

1) *LRU Page Replacement Algorithm:* This algorithm goes well with the principle of locality. A page that has not been used for a long time is least likely to be referenced in the near future. The least recently used (LRU) replaces the page in memory that has been used for the longest period of. This algorithm can be implemented by maintaining the backward distance of each page. Whenever a page is referenced its backward distance is set to zero. Backward distance of other pages is incremented by 1. Replace the page in memory that has not to be used for the longest period of time.

Algorithm:

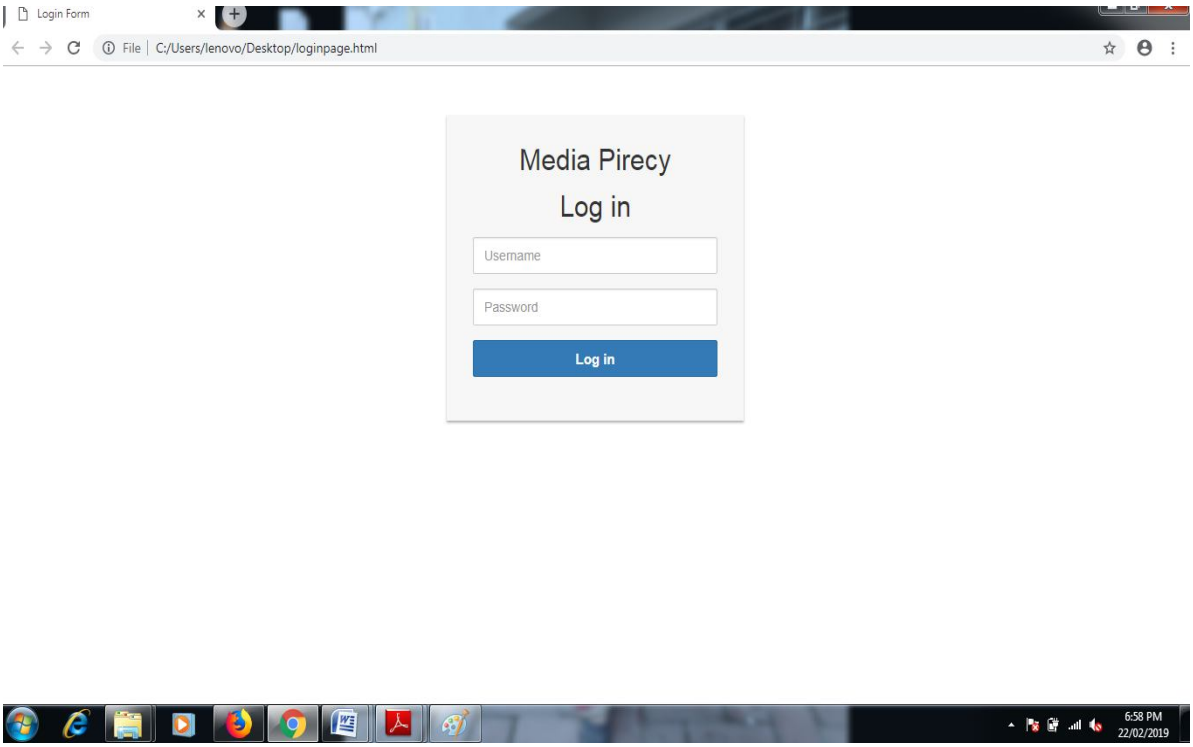
Step 1: Read initial value i.e., number of frames, length of reference string and reference string.
Step 2: initialize array to -1, indicating that the frames are empty.
Step 3: change array size to 0, indicating it will be used for storing backward distance.
Step 4 : For each page reference i in the reference string, if i not in memory and frame=empty then
empty frame=i;
else if
i not in memory and frame!=empty then
longest page distance= i;
i=0;
else if
i in memory then
i=0;
else
i for each page=1;
Step 5: display result.
Step 6: end.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

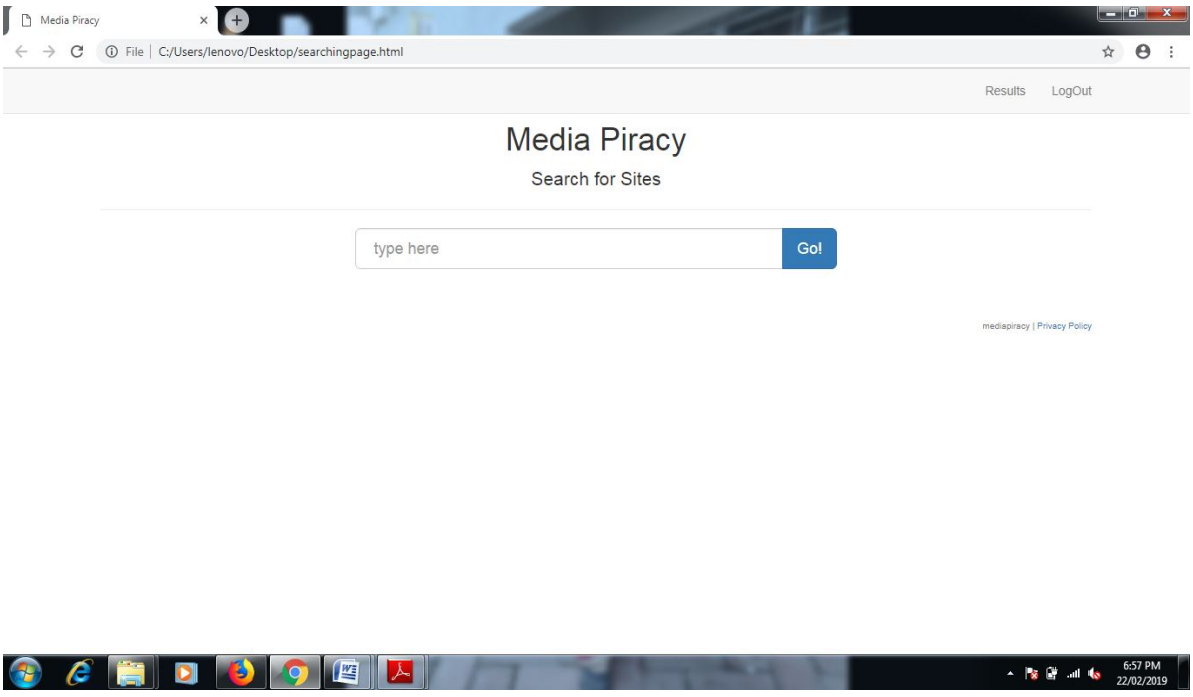
IV. SYSTEM IMPLEMENTATION PLAN

We have divided the implementation plan for proposed system in various set of activities which are needed to be carried out for successful implementation.

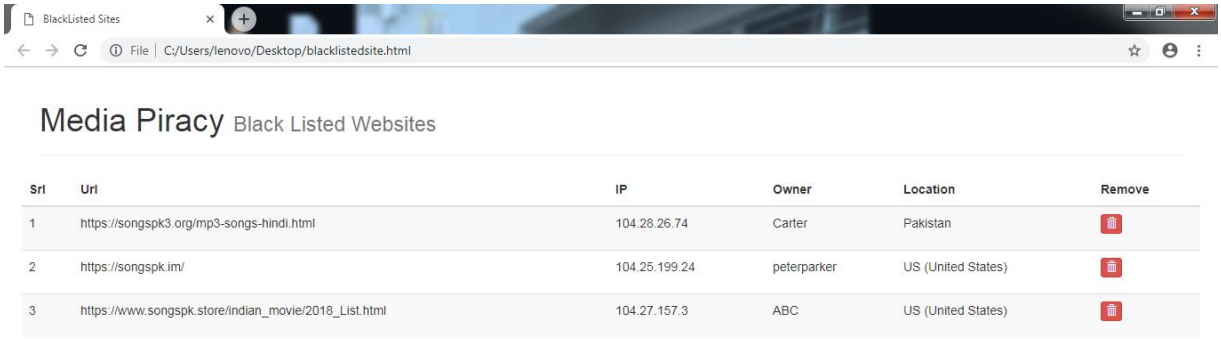
A. Login Page



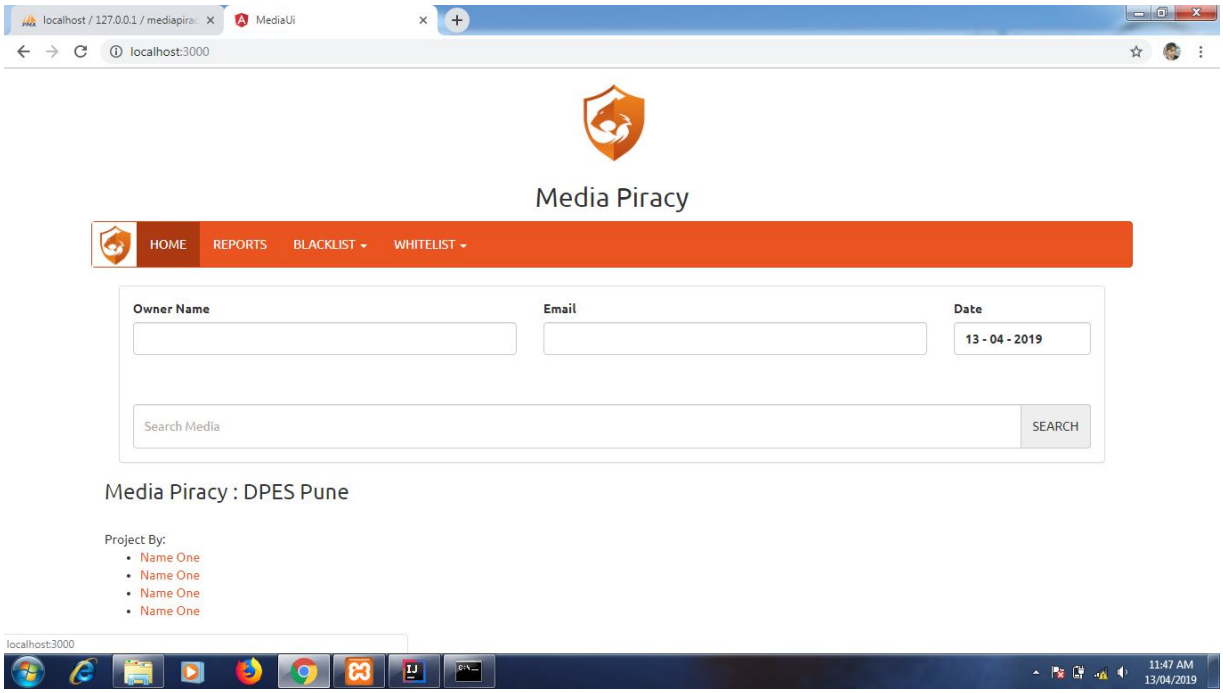
B. Content Searching Page

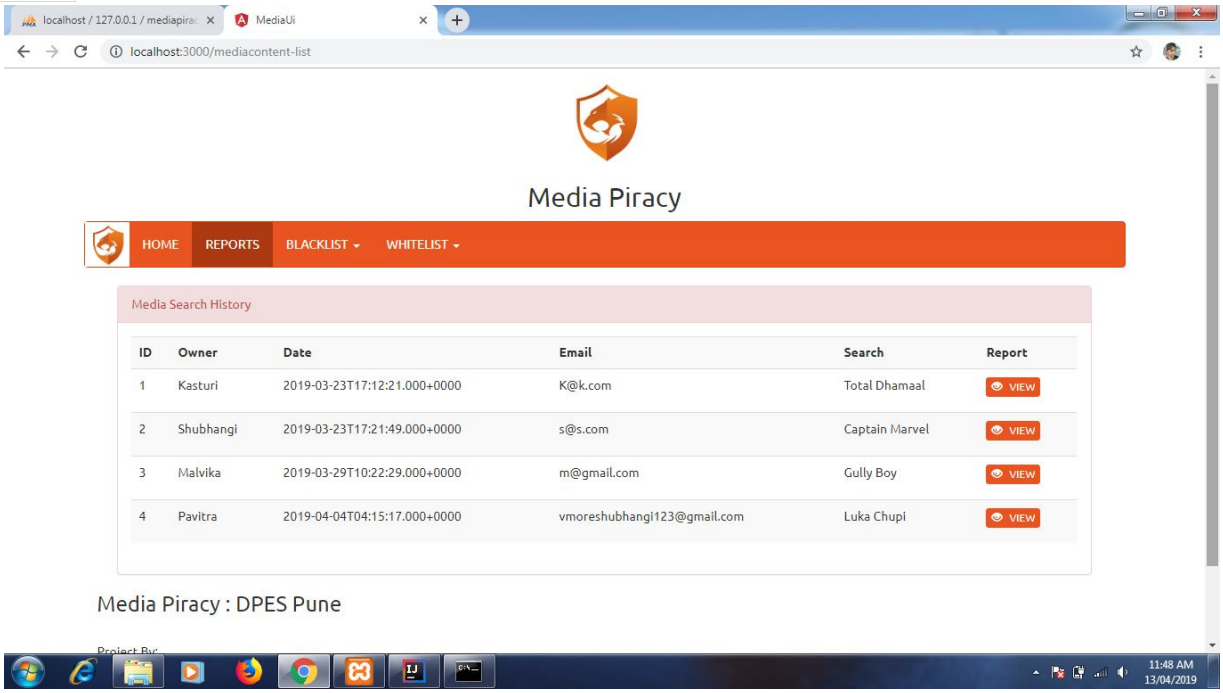


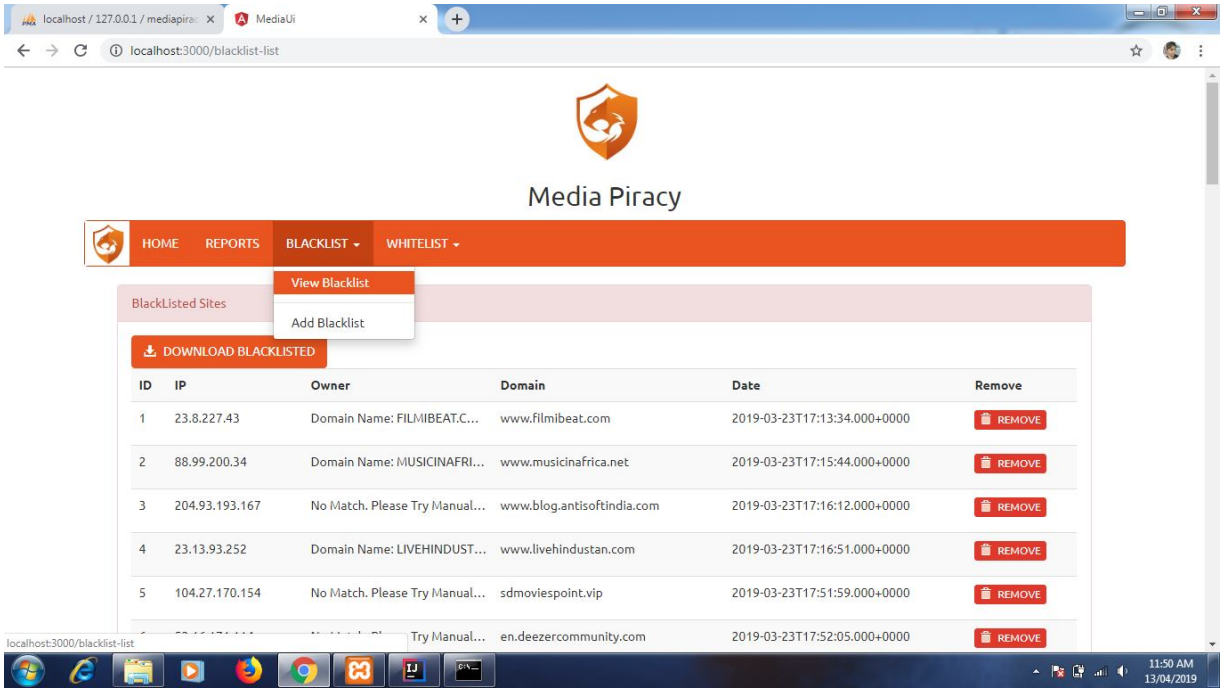
C. Blacklisted contents showing Page



V. RESULTS





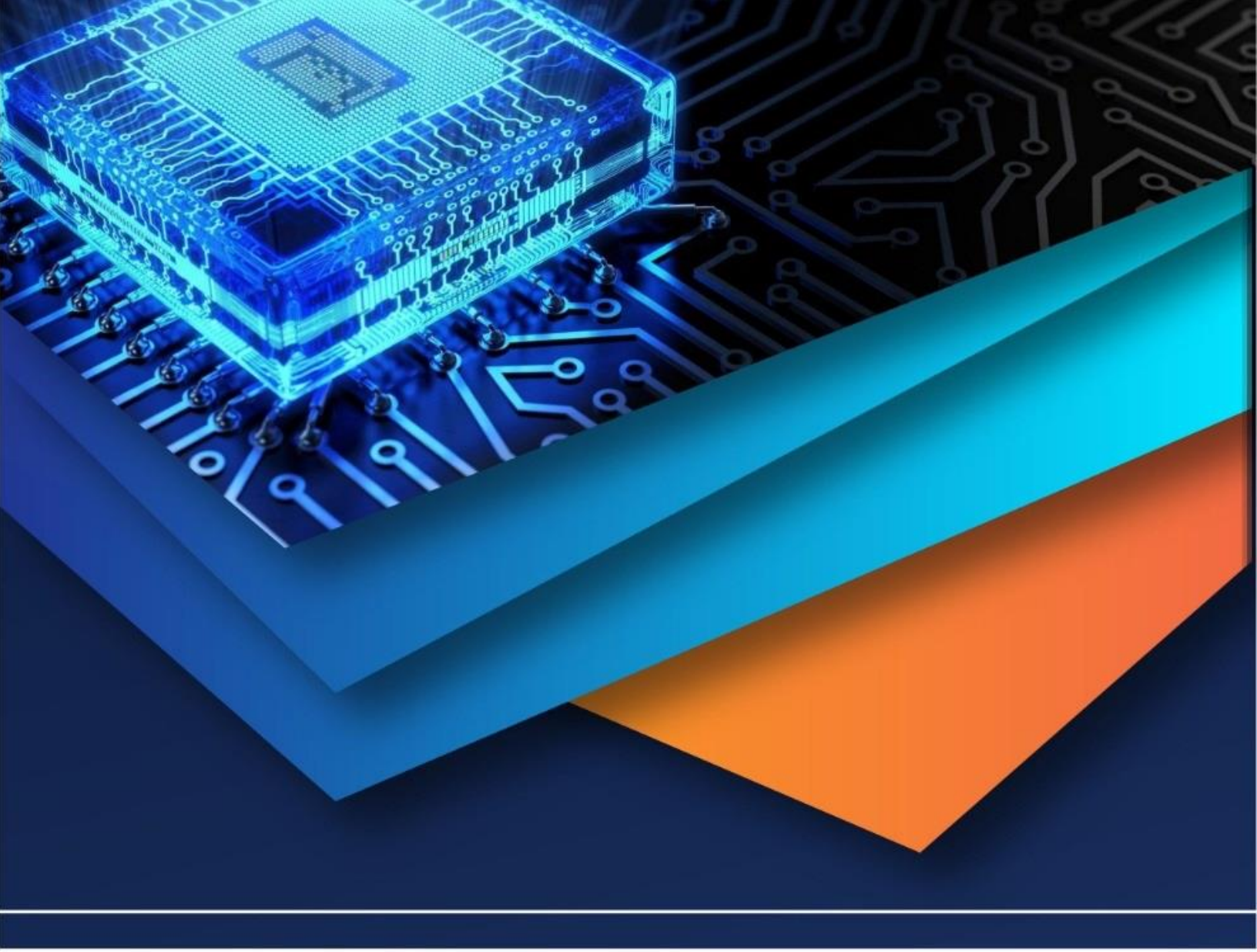


VI. CONCLUSION AND FUTURE WORK

In this System it is presented that how the problem of pirated contents can be solved using AI, ML and DM. It will be applicable in the implementation of large scale content monitoring system that can track and identify illegal distributed content on the web. The system will produce statistical report of media content with information of it such as location, time, IP Address, etc. using different web services and machine learning techniques. Also the system will be storing all the information regarding the blacklisted or the untrusted websites. In future we can use this data to analyse the previous searched blacklisted content and also, we can work on handling huge and large amount of data set.

REFERENCES

[1] Andri Lareida, Burkhard Stiller,"Bit Torrent Measurement: A Country, Network and Content-Centric Analysis of Video Sharing in BitTorrent", IEEE, 978-1-5386-3416-5/18/31.00 c 2018 IEEE 2017.
[2] Tobias HoBfeld,"The Bit Torrent Peer Collector Problem", 978-3-901882-89-0 @2017 1F1P 2017.
[3] Milosh Stolikj, Dmitri Jarnikov, Andrew Wajs,"Artificial Intelligence For Detecting Media Piracy Digital Object Identifier", 0.5594/JMI.2018.2827181 Date of publication: 22 June 2018.
[4] D. Leporini, "Architectures and Protocols Powering Illegal Content Streaming over the Internet", Proc. Int. Broadcasting Convention, p.7, 2015.
[5] By Milosh Stolikj, Dmitri Jarnikov, and Andrew Wajs, "Artificial Intelligence for Detecting Media Piracy", in 2018.
[6] R. Cuevas, N. Laoutaris, X. Yang, G. Siganos, and P. Rodriguez, "Deep Diving into BitTorrent Locality", IEEE INFOCOM 2011, Shanghai,China, April 2011.
[7] M Jain, H. Jegou, and P. Gros. "Asymmetric hamming embedding: taking the best of our bits for large scale image search", In ACM Multimedia, pages 1441–1444, 2011.
[8] AKarpathy et al.,"Large-Scale Video Classification with Convolutional Neural Networks", Proc. IEEE Conf. Comput. Vision Patt.Recogn., pp. 17251732, 2014.
[9] H Jegou and O. Chum." Negative evidences and co- occurences in image retrieval: The benefit of pca and whitening". In ECCV, pages 774–787, 2012.
[10] HJegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search", In ECCV, 2008.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY