A
Project Report
On
"Protectnetic"

Prepared by
TANMAY DAMLE (19DCE021)

Under the guidance of

Prof. Dhruvi Gosai

A Report Submitted to

Charotar University of Science and Technology

for Partial Fulfillment of the Requirements for the

6th Semester Software Group Project-IV (CE359)

Submitted at

CE

DEPSTAR

At: Changa, Dist: Anand – 388421

April 2022

CERTIFICATE

This is to certify that the report entitled "**Protectnetic**" is a bonafied work carried out by **Mr. Tanmay Damle** under the guidance and supervision of **Assistant Prof. Dhruvi Gosai** for the subject **CE359-Software Group Project-IV** (CE) of 6th Semester of Bachelor of Technology in **DEPSTAR** at Faculty of Technology & Engineering – CHARUSAT, Gujarat.

To the best of my knowledge and belief, this work embodies the work of candidate himself, has duly been completed, and fulfills the requirement of the ordinance relating to the B.Tech. Degree of the University and is up to the standard in respect of content, presentation and language for being referred to the examiner.

Ms. Dhruvi Gosai
Assistant Professor
CE DEPSTAR, Changa, Gujarat.

Dr. Dweepna Garg
Head of Department - Computer Engineering,
DEPSTAR CHARUSAT, Changa, Gujarat.

Dr. Amit Ganatra Principal,
DEPSTAR Dean, FTE
CHARUSAT, Changa, Gujarat.

DECLARATION BY THE CANDIDATE

I hereby declare that the project report entitled **"Protectnetic"** submitted to **Devang Patel Institute of Advance Technology and Research (DEPSTAR), Changa** in partial fulfilment of the requirement for the award of the degree of **B.Tech** in Computer Engineering, from **Devang Patel Institute of Advance Technology and Research (DEPSTAR)**, is a record of bonafide CE359 - Software Group Project - IV (project work) carried out by us under the guidance of **Prof. Dhruvi Gosai**. We further declare that the work carried out and documented in this project report has not been submitted anywhere else either in part or in full and it is the original work, for the award of any other degree or diploma in this institute or any other institute or university.

ABSTRACT

Child Abuse is a major concern in the society today. Many people on the dark web or even the surface web host child abusive content on the internet. Since it's not feasible for humans to identify 100s of 1000s of links every day; we made a website which can identify child abusive content with just a click of a button. Our goal - to make a website for scraping the dark web website, analyse the content in it and generate a report for the content within it & throw an alert if child abusive content is found on the provided dark web website.We Believe that the data collected with this project will have various social & business impacts on the society.

ACKNOWLEGEMENT

I would like to express my special thanks of gratitude to my guide **Prof. Dhruvi Gosai** who gave us the opportunity to do this project on the topic **"Protectnetic"**, and was made herself available for all the queries, questions and suggestions from her busy schedule We are really thankful to her for the same.

I would also like to express my thanks to **Prof. Mohammed Bohra (DEPSTAR), Prof. Mrudang Patel (CSPIT), Mr. Dhaiwat Mehta (Dark web expert) for** making themselves available whenever needed and for giving valuable suggestions & support which benefited the team for having a smooth workflow & saved us a lot of time.

TABLE OF CONTENTS

TABLE OF FIGURES

1. INTRODUCTION

1.1 Origin of Idea

Child abuse has become a major problem in our society today. Most people on the dark web or surface web host are child abuse content online. As it is not possible for people to recognize 100 of 1000 links per day. we have created a Web site that can identify the content of child abuse by clicking on a button. To make a website scratch a dark website, analyses the content of the website and produce a report on its content and drop a warning if the content on child abuse is available on a given dark site.

This idea was originated by the team member Jeet Undaviya (19DCE142) while brainstorming and the team came to a conclusion for making this project, as something like this isn't made before in practical.

Team members include:

- Tanmay Damle (19DCE021)
- Yash Kavaiya (19DCE056)
- Jeet Undaviya (19DCE142)
- Raj Soni (D20DCE176)

1.2 What is Protectnetic?

In today's world internet has become available to almost everyone. This certainly has it's benefits & it's cons too. Nowadays, kids also have mobile-phone / tablet devices on which they can surf dark web, some pornographic dark web websites might be harmful for kids as they might do some permanent trauma to kids. Also on some dark web websites, kids are lured in to brainwash. Because of all of this we decided to make "Protectnetic" so that the data collected through this project can be used in doing something good along with making child-web surfing a little better.

Protectnetic is a website which takes a surface web links /onion/ dark web links or a text file of multiple links or csv file of multiple links as input and in turn generates a report which detects whether the provided link/links has child abusive content in it or not.

1.3 Background Work

1.3.1 Dark Web

Dark web is a place where content is not indexed by search engine but it is accessed by specific software or authorization to access.Dark web's content is a part of the internet which is accessible only to particular browser or through specific network configurations.Dark web is used for keeping internet's activity anonymous or private.But dark side of it is, uses of illegal activity on internet.Mainly it is used by the United States Department of Defense to communicate anonymously,buy now it has become a center point for users wishing to remain anonymous around the world.Dark web uses a technology called "onion routing," which protects users from surveillance and tracking through a random path of encrypted servers.When users uses a site through TOR, their information is routed through lots of relay points like thousands of so that it cover the user's tracks and make their browsing for impossible to trace.

1.3.2 NLP

Natural Language Processing (NLP) strives to build a machine that understands and gave respond to text or voice data and respond which is given by machine is similar to human respond.In NLP, Computer can capable of "understanding" the contents of documentations, including the contextual nuances of the language within them. NLP can then accurately extract data and insights contained in the documents as well as categorize and organize the documents automatically. NLP uses AI to take real-world input, then process it, and make sense of it in a way a computer can understand. Just as humans, Machine have programs to read and microphones to collect audio. And just as humans have a brain to process that input, Machine have a program to process their respective inputs. At some point in processing, the input is converted to code that the Machine can understand.

2. SOFTWARE AND HARDWARE

2.1 Software

2.1.1 Python

This project uses python programming language as the main language of the project & all the implementation part.

Figure 1 Python

2.1.2 Django

This project uses Django, a back-end framework for the easement of making the website easily & utilize its different features.

Figure 2 Django

2.1.3 Tor

This project uses tor project for scraping the surface links / onion links.

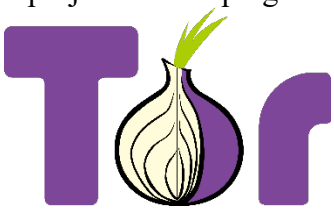


Figure 3 Tor

2.1.4 Front-End

This project uses various front-end technologies for designing.

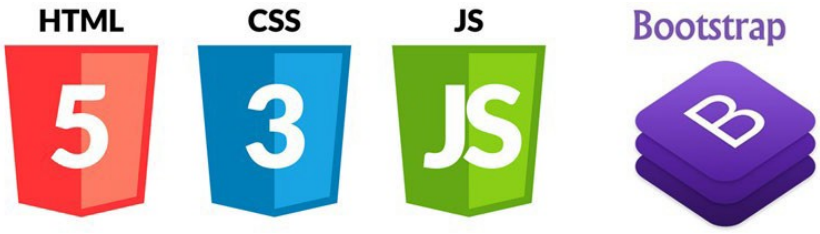


Figure 4 Front-End Technologies

2.1.5 Open CV and TensorFlow

This project uses open cv & TensorFlow for child-abuse detection image models.

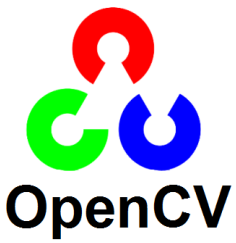


Figure 5 Open CV

Figure 6 TensorFlow

2.1.6 Other Modules

Along with above mentioned software, the project used many other modules for various purposes.

Here is a full list of the modules we use:
https://github.com/damletanmay/Protectnetic-child-abuse-prevention/blob/main/requirments.txt

Figure 7 Celery

Figure 8 Beautiful Soup

2.2 Hardware

2.2.1 Minimum Hardware Requirements

- A system with minimum 4GBs of RAM
- System must contain and install properly all the software
- System might be either Linux, windows or any other Debian based system
- Intel i5 / AMD Ryzen 5/7 to run the project smoothly

3. IMPLEMENTATION & FUNCTIONALITIES

3.1 Implementation

3.1.1 The Beginning

The beginning of this project was like any other project, pretty basic stuff, which includes wireframing, role division, brainstorming ideas to make the website efficiently, estimation of project completion, etc.

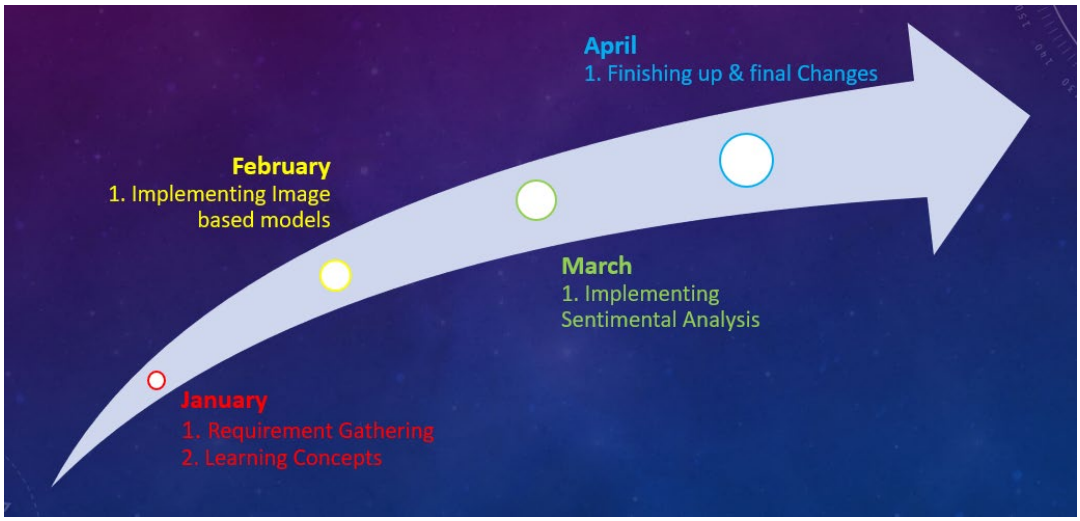


Figure 9 Gantt Chart

3.1.2 Flow of website

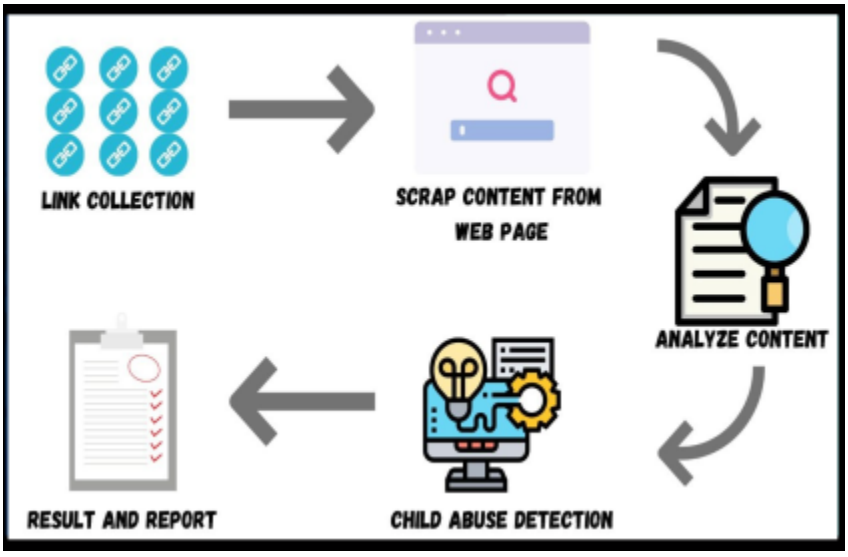


Figure 10 Flowchart

The Work flow is shown in above chart. So, when the user accesses the website, he can enter link or either text file of links or csv file of links to get the results. When the user enters the link, data is scrapped from that link &it goes to the NLP model to determine if child abuse is found or not, if NLP model fails, then all the scrapped images will go to the image processing model & report is saved in the Database & is also shown to the user, if the user wants, he/she/they/them can download the report.

As shown in the figure, the system implementation starts with the link detection which scraps the contents from the web page and analyzing it. If system detects the child abusive contents, then system reports it and accordingly results are generated.
Most of the Child Abusive content is found on dark web so, we have only focused on the detection of the child abuse content from the surface web and the dark web websites. The most famous and largest known dark and deep network is offered by TOR network.

Hence, we firstly managed a way to scrape HTML using the Python requests module from the dark web websites by connecting to the TOR proxy for connecting to the TOR network. We have also used FakeUserAgent for modifying the default headers while sending the GET request to prevent the chances of getting blocked by the website.

Moreover, we used BeautifulSoup Python module with lxml parser to parse the HTML from the scraped response. Then we get all the text found on the website by the get_text() method offered by the module itself. We use this fetched text to detect child abuse content using our NLP Classification Model.

In case our NLP model fail to detect any child abuse on the website based on text then we go for child abusive detection through images and Image Classification Model, which we have designed.

The images on the website are scrapped by <img> tags.The src attribute in img tag contains the link for the image location hosted on the website server.

Then we download the images from the links found from the src attribute in img tag.

And send this downloaded image for the next step are send for child abuse detected by our Image Classification Model.

Find more information about the project on:

https://github.com/damletanmay/Protectnetic-child-abuse-prevention/blob/main/README.md

3.2 Functionalities

- Surface web link processing
- Dark web link processing
- Multiple link processing at the same time
- Database collection for future enhancements
- 75% accuracy with NLP model & 85% accuracy with image processing model i.e., less chances of false positive results.

4. SCREENSHOTS/ OUTPUT

4.1 Screenshots

Figure 11 Home Page

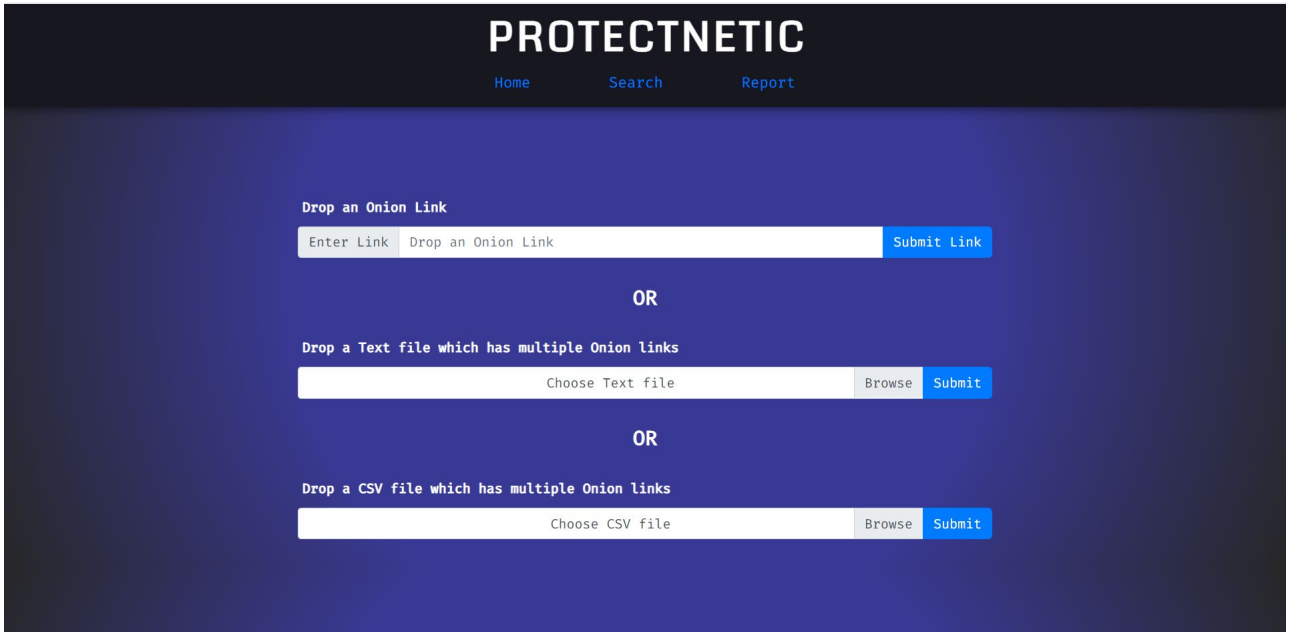


Figure 12 Search Page

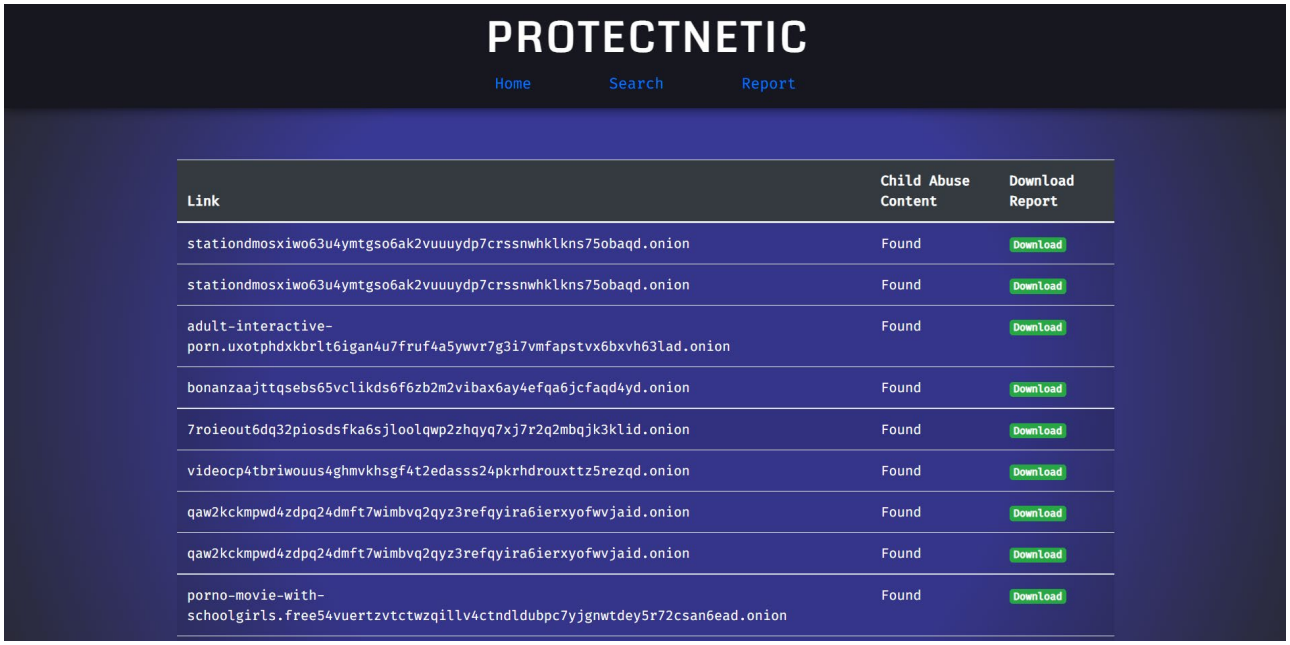
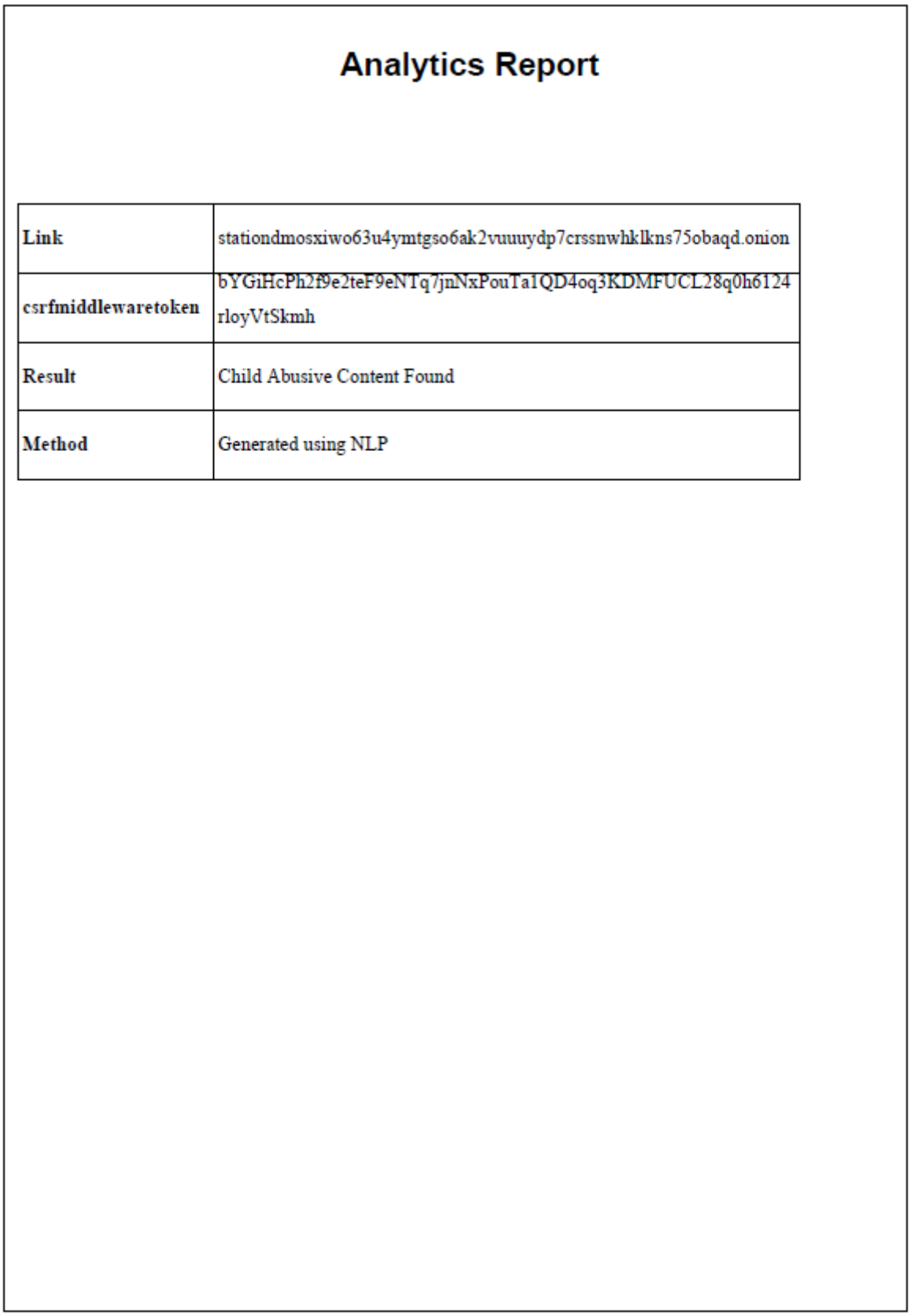


Figure 13 Reports Page

Figure 14 Output



Analytics Report

Link	stationdmosxiwo63u4ymtgso6ak2vuuuydp7crssnwhklkns75obaqd.onion
csrfmiddlewaretoken	bYGiHcPh2f9e2teF9eNTq7jnNxPouTa1QD4oq3KDMFUCL28q0h6124 rloyVtSkmh
Result	Child Abusive Content Found
Method	Generated using NLP

Figure 15 Analytics Report

5. LIMITATIONS

5.1 Limitations

- The website only detects child-abusive websites but it can’t actually close the website permanently
- NLP model sometimes randomly misbehaves due to lack of data whilst training
- Sometimes due to load the website might work slow
- Image processing models are trained well but they take time sometimes if image has noise in it.
- Uses extra server which might cause optimization/ hosting problems in the future.

6. FUTURE ENHANCEMENTS

6.1 Future Enhancements

- Collection of Data for various other applications like making security web filter for public and private organizations.
- Face Detection to identify child and/or child-abusers
- To get IP address of tor websites in order to permanently close them
- Spread awareness about child abuse
- The data can be made available to govt. organizations to take further steps in banning of such sites in area/state/region/country etc.

7. REFERENCES

7.1 References

Here mentioned are some of the references to understand technologies used in the project.

- https://docs.python.org/3/
- https://docs.djangoproject.com/en/4.0/
- https://docs.opencv.org/4.x/d9/df8/tutorial_root.html
- https://www.tensorflow.org/api_docs/
- https://beautiful-soup-4.readthedocs.io/en/latest/
- https://youtu.be/iGWbqhdjf2s
- https://youtu.be/o2BYWYS73G0
- https://getbootstrap.com/docs/5.1/getting-started/introduction/
- https://github.com/
- https://www.torproject.org/
- https://pypi.org/project/django-celery/
- https://redis.io/