NET-DPI: Network Filter Using Deep Packet Inspection And Machine Learning Techniques.

by

Ahmed El-Rawy, Mareine Nassef, Mariam Mohie, Marwan Atef

A dissertation submitted in partial fulfillment of the
requirements for the degree of
Bachelor of Computer Science

in

Department of Computer Science

in the

Faculty of Computer Science
of the
Misr International University, EGYPT

Thesis advisor:
Dr. Ghada Khoriba
Dr. Mohamed Sobh
Eng. Silvia Soliman

(**June 2018**)

Abstract

NET-DPI is an internet filter for an organization's network. It filters the organization's network traffic to achieve network resource optimization and a controlled environment set by its administrators. Most internet filters block entire websites if some or all of the website's content is deemed harmful/irrelevant; however, that means also blocking potentially beneficial content as well by default. The contribution discussed in this paper, is the achievement of accuracy and resource utilization by combining multiple techniques from different domains to solve this problem. This is done by reaching into more depths to access each web page and judging it on its own merits, instead of only taking action according to the website as a whole, or depending on the website's description. This paper explains the main steps to achieve these objectives: intercepting the network flow using firewall (with SquidGuard); extracting packets using Deep Packet Inspection (with tcpdump); then analyzing the content of the web page using a combination of machine learning and deep learning classifiers(K-Nearest Neighbor, Gradient Boosting, and Recurring Neural Network); and accordingly, reaching a decision to allow or block that web page.

Acknowledgments

Firstly, We would like to express our sincere gratitude to our advisors Prof. Ghada Khoriba, Eng. Silvia Soliman, and Eng. Youssef Mobarak for their continuous support of our project, their patience, motivation, and immense knowledge. Their guidance helped us in all the time of research and writing of this thesis. Besides my advisors, we would like to thank the rest of our graduation project committee: Prof. Ayman Ezzat, Prof. Ayman Bahaa, Prof. Ashraf Abdel-Raouf, Prof. Ayman Nabil, and Prof. Eslam Amer, for their insightful comments and encouragement, but also for the hard questions which pushed us to widen our research from various perspectives. Last but not least, we would like to thank our families for supporting us throughout writing this thesis and our lives in general.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

1.1.1 Background

Firewall and network traffic monitoring are quite important nowadays whether in small companies, large companies, schools, universities, and also homes. Internet Traffic monitoring is the process of monitoring all incoming and outgoing data from the internet to a device, network or environment for all the purpose of administration and detecting any abnormalities. Deep Packet Inspection (DPI) is the process of looking beyond the packet header information and extracting information based on the content. DPI is context-aware as well; meaning a packet out of millions would not make much sense, but the concept of DPI is to see all the other packets, concatenating them together to produce the content of any transmission. The data from various sources is then gathered, reviewed and then analyzed to reach a conclusion. Our proposed solution focuses on collecting packets flow in network and make some analysis and classification on it to see whether it needs to be blocked or allowed.

1.1.2 Motivation

This project aims to help companies, schools, universities, and even parents to block all the inappropriate URLs, and allow only URLs that are beneficial to them. Some of the questions that were asked in the survey is that which websites in their working environment is blocked. Majority of the answers were Facebook and YouTube. This project aims to

solve this problem, why block an entire website, when you could just block the URLs that dont meet the requirements of the user.

According to the survey we have made, which targeted system administrators and IT technical support employees in different organizations, stated that 75% uses network filtering and monitoring. 68.75 preferred deep packet inspection over traditional packet inspection. So that if they had a system that does not totally blocks unwanted websites but allow only useful input (like tutorials on YouTube), do they find this system useful enough? 87.5% answered yes while 12.50% answered no. One of the answers is that they did not find a system like this.

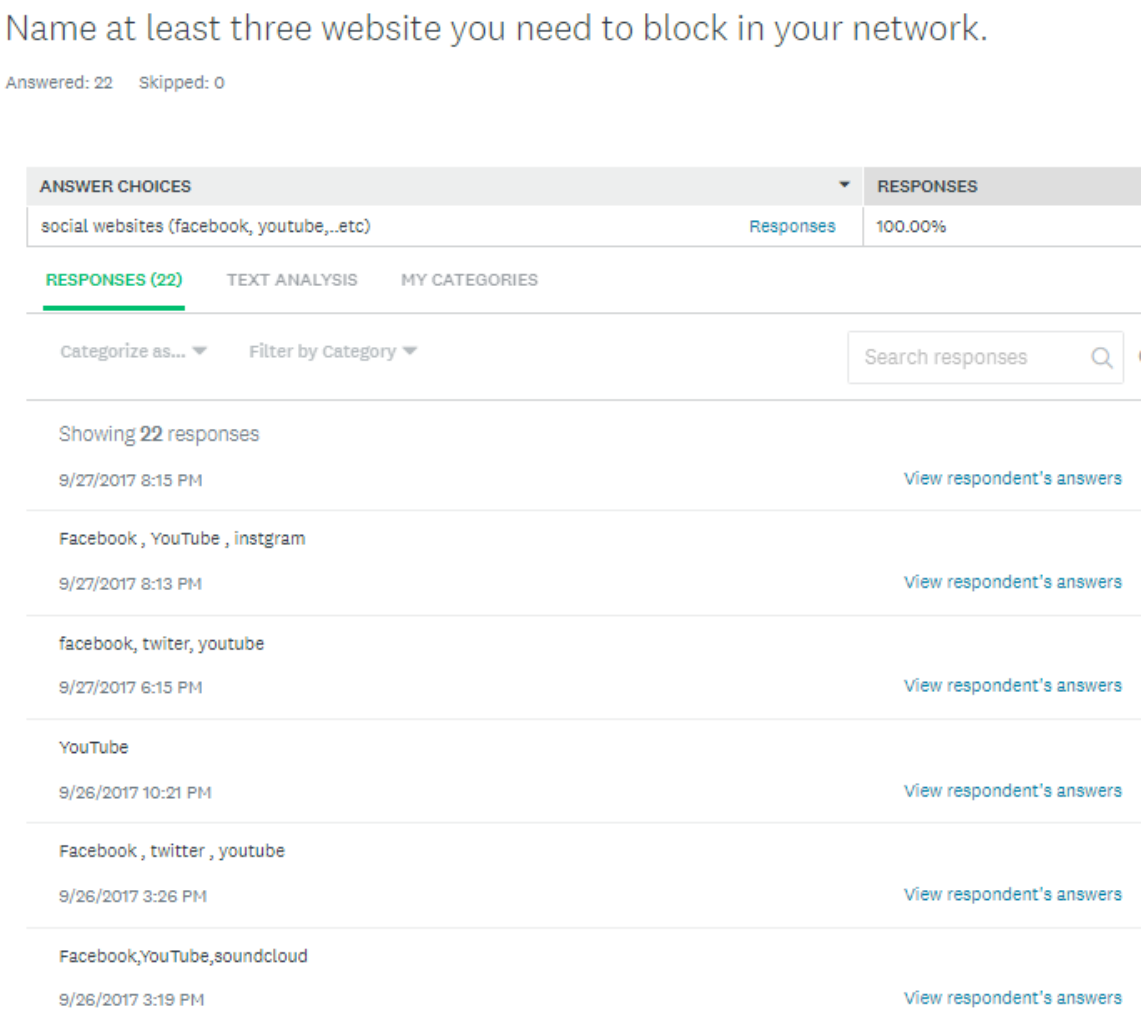


Figure 1.1: survey results

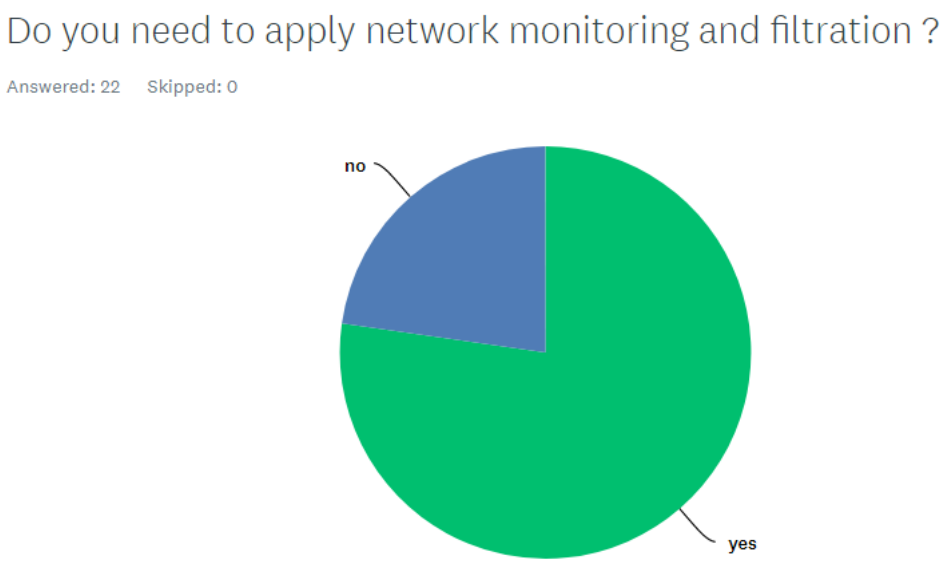


Figure 1.2: survey questions

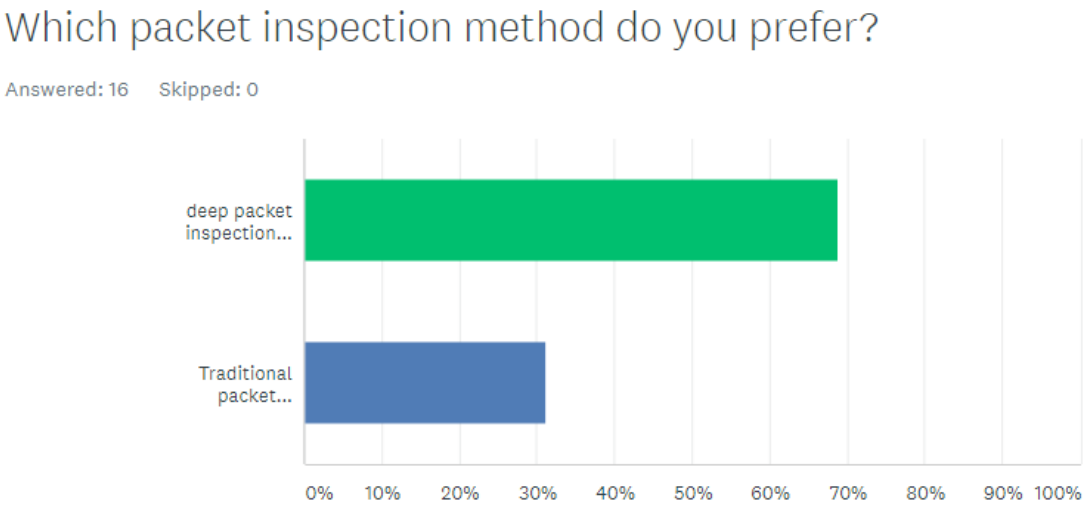


Figure 1.3: survey results 1

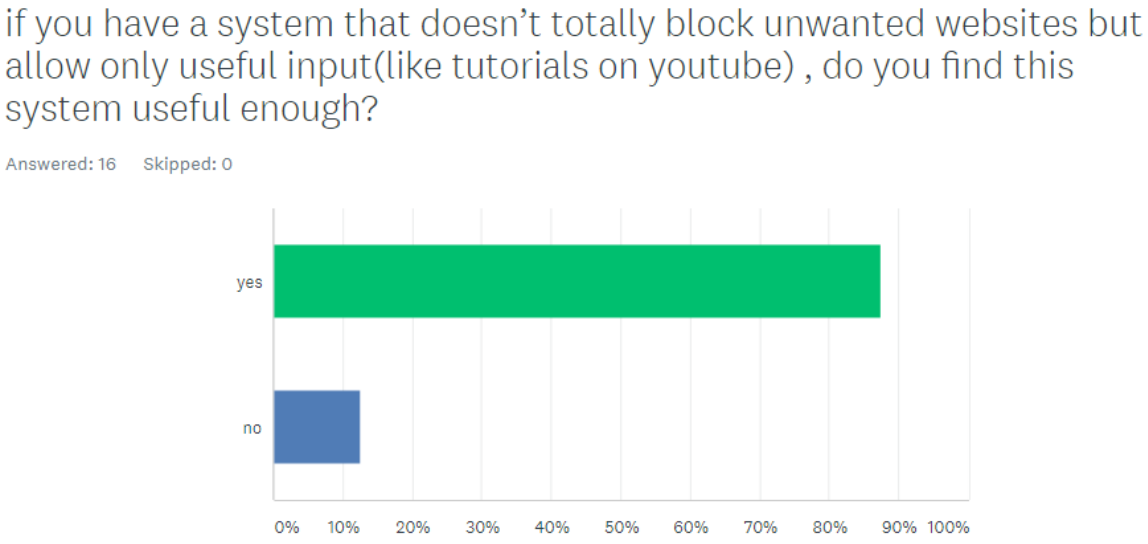


Figure 1.4: survey results 2

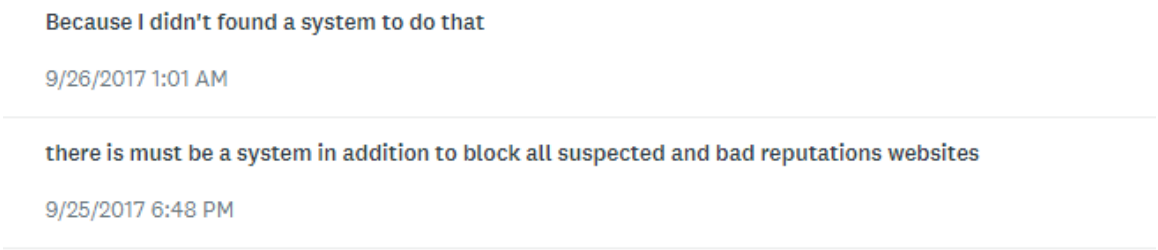


Figure 1.5: survey result why no, to the previous figure 1.4

1.1.3 Problem Definitions

Most of the of companies block websites like Facebook and YouTube to be used by their employees. The goal is allow only certain aspects from these websites to be used in a working environment to benefit the employees in their work, since some of these websites like Facebook and YouTube are important to use like tutorials on YouTube. And the way we are going to obtain this information from packets by using a DPI tool. Also allow a more precise analysis to whatever the input will be. As for the intended users, it shall range from an owner of a company with internal network to governments. Now, the system goal is to monitor network traffic and blocking any unknown or unwanted traffic to prevent malware attack, or misuse of company network resources. Without the traditional blocking mechanism of blocking the entire web application and servers, but intelligently restricting and allowing access to certain ones according to what the clients system needs.

1.2 Project Description

Monitoring and filtering the network, block inappropriate websites and allow specific useful websites.

1.2.1 Objective

The system is designed to work with smart Access. It is something needed constantly for some people, especially those who work mainly on networks field. The purpose of Smart

Access is to allow access to applications that their content is considered beneficial and prevent those that contain irrelevant and sometimes inappropriate content. The kind of softwares that is currently available in the market is that they block websites and protocols as a whole, without putting its content in consideration. What this system ultimately do is that accessing the content of countless packets, analyze it and then classify their contents, and finally according to some analysis and classifications. At the end the system decides whether to allow or restrict access to the requested content.

1.2.2 Scope

1. Owners can monitor the users' network flow.

2. Owners can allow or block specific URLs, not the entire domain.

3. Give the subordinates access to useful URLs only.

4. The System can work on large companies small companies, schools, and homes.
5.

1.2.3 System Overview

In figure 1.6, a wide overview is shown of the main components that make up NET-DPI's methodology and the flow between them, as well as the input and output of each phase. As previously explained, NET-DPI is the combination of all of these components together. The way these components are used is the contributional aspect of this methodology. The network flow of the organization runs through the DPI filter, and if the desired traffic is detected (online video streaming website reply), the rest of the components start working. The firewall then stops the traffic from reaching its destination (a client using the network within the organization). Consequentially, the output of the DPI is then taken as input for the data analysis and Classification phases. The data is analyzed, then a decision is reached, and finally an action is taken by the firewall accordingly.

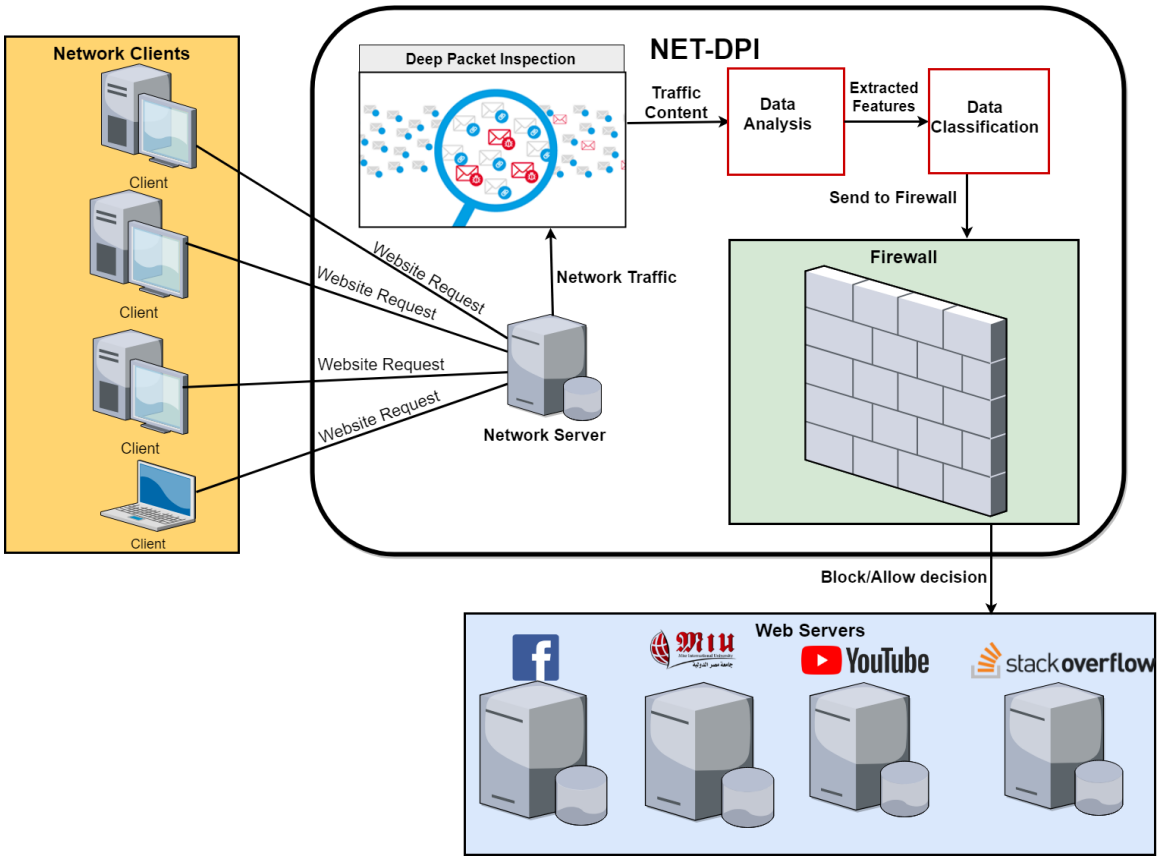


Figure 1.6: System Overview

1.2.3.1 Deep Packet Inspection

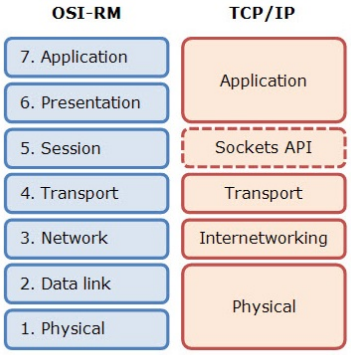


Figure 1.7: The OSI and TCP/IP Layered Architectural Models of Modern Networking [3]

Packets travel across the network layer of the Transmission Control Protocol/Internet Protocol (TCP/IP) layer[5] (figure 1.7). As shown in Figure 1.8, Packets carry certain information that make it reach its destination; whether the sender's IP address, the intended receiver's IP address, something that tells the network how many packets has been broken into and the number of this particular packet as well as the data of the packets (payload). There are two ways to read these packets: Packet Inspection and Deep Packet Inspection. Packet Inspection [6] is the reading of the headers at the network layer, as shown in figure 1.7, to know information about the data (such as destination, source, size,..), without reaching the payload at all. Deep Packet Inspection (DPI) means reaching the packet through multiple layers, not just the network layer. Not only does this mean getting the payload, but also getting it in multiple forms: binary (from Physical Access Layer), ASCII(from Network and Transport Layers), or in the data's original form (Application Layer). DPI can be used for protocol detection, anti-virus, anti-malware and Intrusion Detection System (IDS) [7]. The purpose of DPI in this paper is protocol detection and data extraction. DPI is used to get the payloads of all packets of a certain internet transaction (one web page) and turning it to their original form: web page source (i.e. text). In conclusion, the DPI tool will inspect all the network flow of the organization and if any packets are found to be from the online streaming video website, then the firewall is called to pause these packets from reaching their destination until computations are made and a decision is taken.

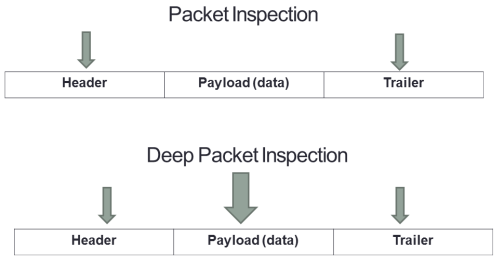


Figure 1.8: The Difference between Standard and Deep Packet Inspection

1.2.3.2 Firewall

In computing, a firewall is a network security system that monitors and controls incoming and outgoing network traffic based on predetermined rules[8]. For NET-DPI, the rule is to allow any traffic except the rejected content categories provided by the organization administrators after applying it to the data analysis phases. For example, if the organization is a university and its Dean wanted only education videos to be allowed, then it is the firewall's task to block unwanted traffic of all other categories and allow the wanted educational traffic. However, the firewall cannot know the type of websites allowed or not; that is the data analysis phases' task. The type of firewall used for this system is called a "network firewall", also known sometimes as a "packet filter". Packet filters look at network addresses and ports of packets to determine if they must be allowed or blocked[9]. The firewall tool used for this system gets the traffic's URL from the DPI component, then pauses it from reaching the user whom had requested it. The time during the pause, the analysis phases are run over the page source outputted by the DPI component as well. When the analysis and classification are done, a decision is given to the firewall to allow or block that URL. The action of allowing or blocking is in fact redirecting to another URL. The firewall, resides as a part of the network server, and controls clients' traffic by redirecting web page request traffic to other kinds of traffic for example a block page.

1.2.4 Data

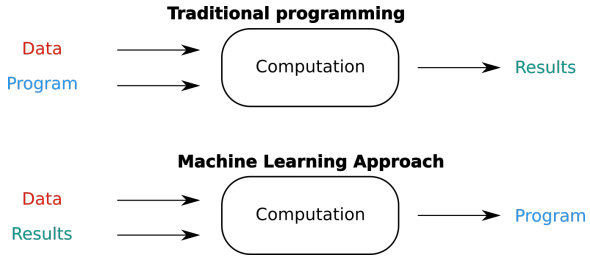


Figure 1.9: The Difference between Traditional Programming and Machine Learning Approach [4]

The data contained in any video streaming page is either text or video, and for both, classification and analysis are used so that the system could "understand" the content. As shown in figure 1.9, Machine Learning is an approach used to find patterns and similarities and extract rules to follow. Classification is a part of the Machine Learning approach, and that part is called supervised learning. Classification is a training set of correctly identified observations given, so when the input is unidentified, the machine can still categorize it[10]. Deep Learning or Deep Neural is an advanced approach to Machine Learning, designed to act as an artificial Neural Network copying that of the human brain [11]. Both Machine Learning and Deep Learning classifiers use features extracted from the original content. However, Deep Learning classification algorithms can function without feature extraction, so they can work with data such as text, hexadecimal, or bytes, as well as encrypted content.

The main type of data that NET-DPI uses for web page content analysis is text. Text is usually unstructured, which means it cannot be the input to uniform computation. A web page source, which is the script that makes the web page appear as it does, is semi-structured because of its tags. However, for classification algorithms to use this text as input, it must be analyzed and structured first. That is called Feature Extraction. During the analysis phase, the text undergoes some Natural Language Procedures to convert it from unstructured to structured data, an example of that is shown in figure 1.10. These Procedures are: the removal of any parts of the text that is not part of the English language or any other language detected; then the removal of non-essential words that do not contribute to a sentence's meaning (such as: "the", "a", "an"), and stemming, which is returning all words to their respective origins ("entertainment", "entertaining", and "entertained" all become

"entertain"); finally using a pre-prepared bag of words for each category to match the input and turn the unstructured text to structured numerical data. The categories below act as both features and classes.

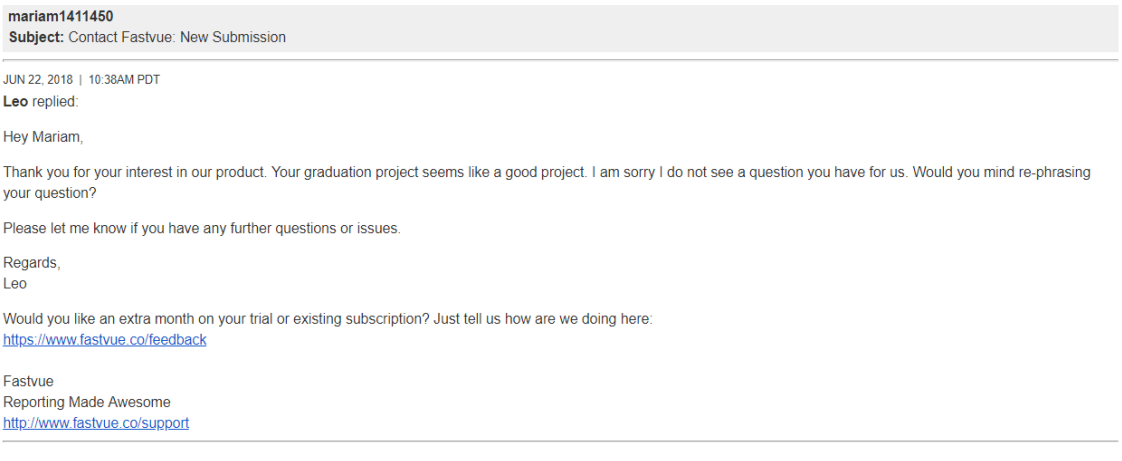
- Film and Animation
- Autos and Vehicles
- Music
- Pets and Animals
- Sports
- Travel and Events
- Gaming
- People and Blogs
- Comedy
- Entertainment
- News and Politics
- How-to and Style
- Education
- Science and Technology
- Nonprofits and Activism

film and animation	education	autos and vehicles	entertainment	class
4	2	3	33	entertainment
17	10	0	55	entertainment
1	6	0	18	entertainment
0	7	0	22	entertainment
0	135	0	18	entertainment
0	0	0	19	entertainment
1	0	0	11	entertainment
2	0	0	16	entertainment
1	6	5	2	entertainment
0	4	0	74	entertainment
9	275	0	57	music
9	268	0	90	music

Figure 1.10: An Example of the Initial and Final states of the Input Text

1.3 Project Management and Deliverable

1.3.1 Supportive Documents



mariam1411450
Subject: Contact Fastvue: New Submission

JUN 22, 2018 | 10:38AM PDT
Leo replied:

Hey Mariam,

Thank you for your interest in our product. Your graduation project seems like a good project. I am sorry I do not see a question you have for us. Would you mind re-phrasing your question?

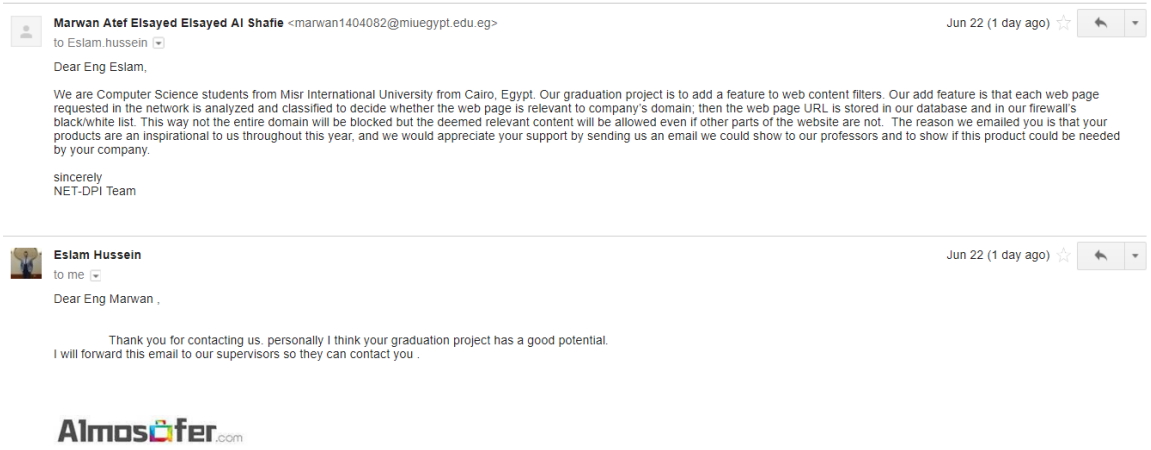
Please let me know if you have any further questions or issues.

Regards,
Leo

Would you like an extra month on your trial or existing subscription? Just tell us how are we doing here:
https://www.fastvue.co/feedback

Fastvue
Reporting Made Awesome
http://www.fastvue.co/support

Figure 1.11: Sponsor 1 fastvue



Marwan Atef Elsayed Elsayed Al Shafie <marwan1404082@miuegypt.edu.eg> Jun 22 (1 day ago)
to Eslam.hussein

Dear Eng Eslam,

We are Computer Science students from Misr International University from Cairo, Egypt. Our graduation project is to add a feature to web content filters. Our add feature is that each web page requested in the network is analyzed and classified to decide whether the web page is relevant to company's domain; then the web page URL is stored in our database and in our firewall's black/white list. This way not the entire domain will be blocked but the deemed relevant content will be allowed even if other parts of the website are not. The reason we emailed you is that your products are an inspirational to us throughout this year, and we would appreciate your support by sending us an email we could show to our professors and to show if this product could be needed by your company.

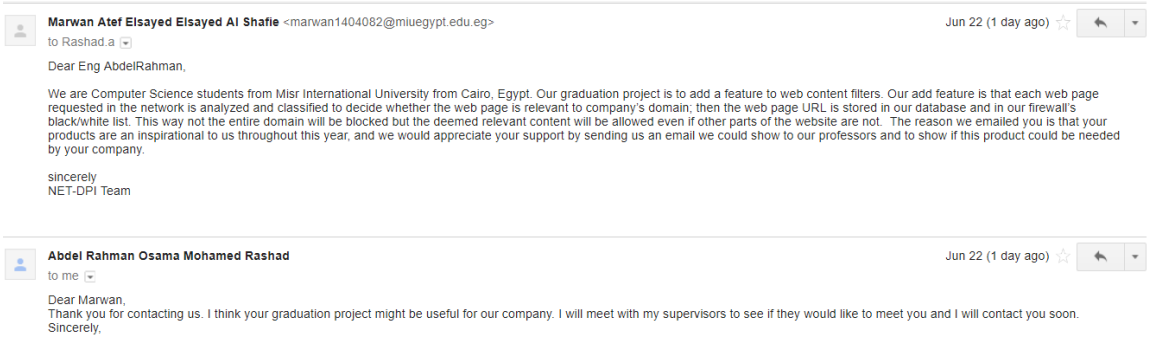
sincerely
NET-DPI Team

Eslam Hussein Jun 22 (1 day ago)
to me

Dear Eng Marwan ,

Thank you for contacting us. personally I think your graduation project has a good potential.
I will forward this email to our supervisors so they can contact you .

Figure 1.12: Sponsor 2 El Mosafer



Marwan Atef Elsayed Elsayed Al Shafie <marwan1404082@miuegypt.edu.eg> Jun 22 (1 day ago)
to Rashad.a

Dear Eng AbdelRahman,

We are Computer Science students from Misr International University from Cairo, Egypt. Our graduation project is to add a feature to web content filters. Our add feature is that each web page requested in the network is analyzed and classified to decide whether the web page is relevant to company's domain; then the web page URL is stored in our database and in our firewall's black/white list. This way not the entire domain will be blocked but the deemed relevant content will be allowed even if other parts of the website are not. The reason we emailed you is that your products are an inspirational to us throughout this year, and we would appreciate your support by sending us an email we could show to our professors and to show if this product could be needed by your company.

sincerely
NET-DPI Team

Abdel Rahman Osama Mohamed Rashad Jun 22 (1 day ago)
to me

Dear Marwan,
Thank you for contacting us. I think your graduation project might be useful for our company. I will meet with my supervisors to see if they would like to meet you and I will contact you soon.
Sincerely,

Figure 1.13: Sponsor 3 GIG Arab Misr Insurance Group

Chapter 2

Literature Work

2.1 Similar System Information

The techniques used in this similar system, and its methodology of using these techniques together, are relevant and considerably new as seen in paper[12]. It mentions the usage of similar techniques to achieve different results. Moreover, it discusses how much network traffic filtering and classification have come to light due to the quick growth of online websites and applications.

Furthermore, another related work [13] is one that advocates how much network filtering and classification is of vital importance for network management and network security. The paper discusses the concept of the majority of unknown traffic is conducted by certain types of applications; it gives this unknown traffic a name: Elephant Traffic. The authors state that traffic sharing the same server IP and the same server port is generated by the same application, and belongs to the same service. Therefore their proposed solution is a novel method, where statistical features are used for cluster analysis, to classify this Elephant Traffic. In order to filter this traffic, nDPI is used, which is an open-source DPI library. It is used to filter through this unknown traffic (Elephant Traffic). In this similar system, the problem their authors trying to solve is the unknown traffic that is being generated by few or some specific kinds of applications (which is referred to as Elephant Traffic), which is considered a different problem from NET-DPI's problem.
In addition to filtering network traffic using deep packet inspection, the corresponding paper [14] demonstrates how certain patterns, because of their reiteration, can cause DPI to slow

down the process of filtering traffic. Furthermore, these repeated patterns can be scanned only once and then skipped if encountered again.
As mentioned in paper [10], automated classification has witnessed a boost in interest. The rapid growth of machine and deep learning has taken on added importance in the last 10 years. The authors of the paper state that the reason for this growth is due to the increased availability of digital documents and the increasing need to organize them. The purpose of their paper is to discuss the main approaches to text categorization and the usage of machine learning in it. Furthermore, in their paper it is demonstrated in details the classifier constraints. Moreover, the paper discusses how Text Classification (TC) is an instance of Text Mining. Furthermore, TC has been applied in several projects whether in document filtering, automated metadata generation. The previous paper provided guideline to the world of text mining and text classification.

Table 2.1: Used algorithms and their accuracies [1]

	Enterprise	Malware
Algorithm	Standard Enhanced	Standard Enhanced
LinReg	99.92% 99.28%	0.00% 58.65%
L2-LogReg	93.35% 98.36%	16.86% 76.13%
L1-LogReg	92.75% 98.97%	19.71% 75.08%
DecTree	97.55% 97.02%	40.98% 83.33%
RandForest	99.53% 99.99%	33.54% 76.79%
SVM	11.94% 99.78%	77.98% 72.62%
MLP	95.90% 99.545	20.61% 72.53%

Worthy of mentioning is another related paper [1] that uses DPI as well as classification algorithms. In their paper, the authors used real network data as an input for six common classification algorithms and their accuracies as shown in table 2.1. Furthermore, these algorithms were not combined together to generate hyper optimized algorithms, on the contrary, each one is tested individually.Even though the previous similar system has the same concept as this thesis, yet what makes NET-DPI exceed it is that after obtaining the best resulted classifier from both deep learning and machine learning classifiers, they are tested one after the other to obtain the ultimate result.

Finally, the authors in this paper [2] advocate that the fact that the internet has become a fast developing environment that is used in every organization (e.g. educational

institution, governments).The authors state that the students have the ability to surf the internet for educational content, but unfortunately they still have the ability to download and surf noneducational content which consumes bandwidth of the network. The majority of the time proxy server maintain a blacklist of Uniform Resource Locators (URL), which are kept as a static list.

2.1.1 Similar System Description

Table 2.2: Used algorithms and their accuracies for first data set [2]

	Word Appearance	Word Frequency
C4.5 (J48 IN WEKA	83.0%	80.5%
Naive Bayes	95.0%	94.5%
PRISM	74.5%	63.0%
Support Vector Machine	95.5%	94.5%

Table 2.3: Used algorithms and their accuracies for second data set [2]

Algorithm	Accuracy	AUC
Naive Bayes Multinomial	92.9%	0.954
SVC	77.5%	0.992
Linear SVC	98.9%	0.993

Combining both DPI with analysis and classification techniques was considered a new approach. The software produced in this paper [12] was considered similar to what NET-DPI is trying to achieve. It demonstrates how network traffic classifying has become a challenge. For that reason, the authors of the paper have proposed a solution, called "Deep Packet": a solution that combines both classification techniques and DPI. Deep Packet classifies network traffic using deep learning. Their classification algorithm is Neural Network (NN), for example CNN with an average accuracy 95%, as well as stacked autoencoder NN (SAE) with its average accuracy 95 %. But what makes NET-DPI exceeds Deeps packet is that the classifications that gave the best results was a combination of both machine learning and deep learning; RNN 92%, CNN 93%, Gradient Boosting 94 %, Decision Tree 95%, KNN 99%.

Furthermore, as mentioned in [14], the authors of this similar system tried to minimize the slowness of dpi by detecting patterns (whether text or bytes) that have been repeated and if encountered again; skip it. Similarly, in this system, the patterns used are of text, obtained from the page source. Even though NET-DPI and the proposed system of the similar system solve the same problem ; yet each ones' approach to the solution is different.

The authors of this related system [2] developed a software that is somewhat similar to NET-DPI. They proposed a solution to use machine learning to predict whether the URL is considered educational or noneducational. As a matter of fact, their algorithms have been tested on two data sets. The first data set had these algorithms: Naive Bayes and two Support Vector Machine (SVM) classifiers, namely SVM with RBF kernel (SVC) as shown in table 2.2 with their accuracies. As for the second data set, the algorithms that were used were: SVM with linear kernel (Linear SVC), Naive Bayes Multinomial SVC as demonstrated in table 2.3. The input of the previous classifiers was the text inside the body tag found in Hypertext Markup Language (HTML) page sources after removing tags and extracting English words. Although the previous system has the same purpose as NET-DPI , yet this similar system purposed solution is not executed in DPI rather than using Squid proxy server.

2.1.2 Comparison with Proposed Project

Points of Compassion	Algorithm(s) used in 1st Dataset	Accuracies (Respectively)	Algorithm(s) used in 2nd Dataset	Accuracies (Respectively)
Deep Packet	Stacked autoencoder (SAE), CNN	95% , 95%	-	-
Accelerated DPI	slow path, data path	-	-	-
Dynamic blacklist of URLs	C4.5 (J48 IN WEKA) , Naive Bayes , PRISM, SVM	Word Appearance: 83% ,95.0% ,74.5%,95.5%	Naive Bayes multinomial, SVC ,Linear SVC	92.9%, 77.7% , 98.9%
NET-DPI	KNN ,Decision Tree , Gradient Boosting	100%, 97%, 99%	KNN ,Decision Tree , Gradient Boosting	99% , 95% , 94%

Figure 2.1: Similar Systems Comparisons.

Chapter 3

System Requirements Specifications

3.1 Introduction

3.1.1 Business Context

There are many types of customers that need the Smart Access feature in their network. Basically any one who wants to filter the subordinates' usage without confining inflexible limitations to the Internet can use NET-DPI system.

There is a need for Network Filters in organizations such as companies or educational institutions; because many of those rely on the Internet, however without the monitoring and filtering of subordinates' Internet usage, that could lead to major waste of resources such as time and productivity which lead to waste of money. Some companies pay enormous amounts of money annually to limit the Internet access for their employees, while other smaller companies cannot afford such functionalities despite their importance. For those organizations that do have one of the commercial systems already in the market, there is a limitation that does not satisfy the employees or students because the whole website is blocked even if it is important sometimes for business need. A simple example for this would be YouTube: Most organizations that do have Network filters tend to block YouTube because of the irrelevant, and sometimes inappropriate content of the application, however when employees or students actually need something relevant to their work on the website,

they cannot access it. This problem can be solved by NET-DPI because it allows Smart Access: a whole website does not need to be blocked, but only block the non-beneficial content.

3.2 General Description

3.2.1 Product Functions

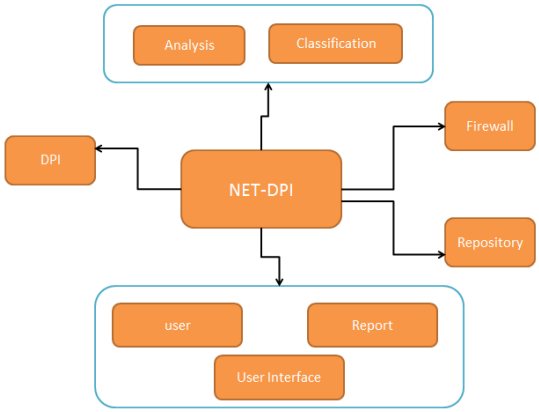


Figure 3.1: NET-DPI Components

each of these modules is a necessity in NET-DPI system.

- User Interface Module: is to allow the user to be able to interact with the interface. but must have advanced knowledge of technology or any knowledge what so ever in computer science.
- Reports: NET-DPI prompts this feature to deliver reports about the User's behavior to the Admin if they see that any web-page is miss classified.
- Deep Packet Inspection (DPI) Module: is responsible for obtaining data from the packets.
- Analysis Module: executed after obtaining the data from the DPI Module to analyze and classify that data.
- Firewall Module: responsible for allowing or blocking the requested services and web-pages according to previous inspection and analysis

3.2.2 User Characteristics

Even though the system has many different potential customers, there are only three types of users: Regular User, Super User, and Admin.

- Admin: Usually the business owner or manager gets this user type. The Admin is the one who decides what types should the other users be. The Admin configures the system how they want it. An Admin needs to have basic computer skills, as well as a complete understanding of the company needs for the system, and is usually the highest one the chain of command.
- Super User: Users of this type are the ones who are in charge of certain departments or divisions. Also the Super User receives reports and notifications about employees under their commands. This type needs basic computer skills, and is in the middle of the chain of command.
- Regular User: This type of user's only interaction with the system is just logging in before using the network. This is the User that is being observed by others who have authority of him/her. This type requires basic computer skills and the knowledge of what is acceptable and what is not in his/her department.

For example: a CEO is the Admin, Department heads are Super Users, and all other employees are Regular Users, with reference to their "parent" Super User.

3.2.3 User Problem Statement

Organizations totally block websites that are deemed irrelevant or inappropriate. But sometimes these websites contain content that is important for business and is work related. An example for this is YouTube: YouTube can have entertainment videos like songs or movies, but can also have videos that are work related such as tutorials. So smart access is needed in this case.

3.2.4 User Objectives

Customer needs are mostly same no matter what is the type of the organization. Customers need smart access to their network that would help them monitor and manage subordinates' behavior, prevent the misuse of resources, as well as minimize access to irrelevant or inappropriate content in the environment. All in an affordable, usable system.

3.2.5 General Constraints

There are some constraints in the system such as the device must be connected to Internet, the operating system deployed must be Linux platform on a Network server.

3.3 Functional Requirements

3.3.1 Class Data

3.3.1.1 ID: f01

-Title: transliterate.
-Description: Removing any special characters
-Input: line of string
-Output: same line in English format
-Pre-condition: words containing special character
-Post-condition: English words
-Dependency: None

3.3.1.2 ID: f02

-Title: Isdigit.
-Description: checking if digit
-Input: word
-Output: true if digit, false if not digit
-Pre-condition: unknown digits
-Post-condition: digits known
-Dependency: None

3.3.1.3 ID: f03

-Title: Start-analysis.

-Description:get the captured file and encode it then remove digits with regular expression. compare each word with the bag of words and then increment the counter of the class.

-Input:captured file from tcpdump

-Output: Row of numbers represents the number of presence of words of specific class in the file

-Pre-condition: Data stored in files as a text

-Post-condition: each file stored as a row in CSV file

-Dependency: Capturing packets and store in file

3.3.1.4 ID: f04

-Title: Start-classification.

-Description: list the dataset in form of features and results

-Input: Row of 16 column

-Output:separate the input file to 15 feature and 1 class

-Pre-condition: data read from the CSV file as a whole (16 column)

-Post-condition: Data read from CSV file as a features and class

-Dependency: Analysis of data file and store the output on CSV file

3.3.1.5 ID: f05

-Title: Call-classifier.

-Description: Pass the row of data in the CSV file to the classifiers (KNN, Decision Tree, Gradient Boosting)

-Input: Row of Data

-Output: Predicting the class of that row of data

-Pre-condition: Unknown class

-Post-condition: Predicted class

-Dependency: Preparing training data before sending it to the classifier by separate features

from the classes.

3.3.2 Class DPI

3.3.2.1 ID: f06

-Title: StartDPI.
-Description: setup the environment whether there will be classification or not depending on the presence of the requested web-page in the list or not.
-Input: captured text file.
-Output: None.
-Pre-condition: .
-Post-condition: taking decision whether to classify or take an action.
-Dependency: Capturing packets.

3.3.2.2 ID: f07

-Title: getURL.
-Description: Open the captured file and search in it for the URL which will help take the decision.
-Input: captured text file.
-Output: string containing the requested URL.
-Pre-condition: presence of the whole text file without knowing the URL to help taking action.
-Post-condition: availability of the URL to be able to take an action.
-Dependency: presence of captured file to search on it.

3.3.2.3 ID: f08

-Title: CheckURL.
-Description: Check if the website is within the system scope or not.
-Input: URL.
-Output: true if within scope, false if out of scope.

-Pre-condition: not knowing if the requested URL is within the project scope or not.
-Post-condition: None.
-Dependency: None.

3.3.2.4 ID: f09

-Title: Cleanhtml.
-Description: Get the captured file and get only the needed data which is the actual data sent from source to destination as there is many unneeded data as handshakes.
-Input: captured file.
-Output: filtered file.
-Pre-condition: file containing so many unneeded data.
-Post-condition: file containg only needed data.
-Dependency: presence of captured data.

3.3.2.5 ID: f10

-Title: GetPageSource.
-Description: Get the page-source of the requested URL and save it into file and encode it.
-Input: URL.
-Output: page-source file.
-Pre-condition: unable to get the content of the packets.
-Post-condition: able to get the content.
-Dependency: run after taking decision that tcpdump cannot get the encoded packets as the requested web-page is https or http and have inner encryption algorithm.

3.3.3 Class FireWall

3.3.3.1 ID: f011

-Title: block.
-Description: blocking certain URL because it classified as non educational.
-Input: URL.

-Output: None.

-Pre-condition: user may browse such URL freely.

-Post-condition: URL blocked by the system.

-Dependency: output of classification.

3.3.3.2 ID: f012

-Title: allow.

-Description: allow user to open such URL as it classified as education.

-Input: URL.

-Output: None.

-Pre-condition: user may browse such URL freely.

-Post-condition: user may browse such URL freely.

-Dependency: output of classification.

3.3.3.3 ID: f013

-Title: CheckList.

-Description: Check the presence of a certain URL in the firewall list.

-Input: URL.

-Output: true if found and false if not.

-Pre-condition: None.

-Post-condition: append the URL if not found and if found block it if needed.

-Dependency: None.

3.3.3.4 ID: f014

-Title: TakeAction.

-Description: save to file URL@True or URL@False and make it happen.

-Input: URL, decision.

-Output: None.

-Pre-condition: user may browse such URL freely.

-Post-condition: user may browse such URL freely. or cannot do it if the action taken is to block that URL
-Dependency: None.

3.4 Interface Requirements

3.4.1 API

- tcpdump
- SquidGaurd

3.5 Performance Requirements

3.5.1 Hardware Interfaces

Routers, Network Interface Cards(NIC), access points, switch and cables and the network server which is core i7, 3.0 GHZ, 4TB hard disk, 16GB Ram.

3.5.2 Communications Interfaces

Routers, Network Interface Cards, and switches.

3.5.3 Software Interfaces

languages and tools: Python, bash, tcpdump and SquidGaurd.

3.6 Other non-functional attributes

3.6.1 Resource Utilization

The Deep Packet Inspection does not run all the time, instead the system firstly deploys standard inspection, which is simpler and less costly in terms of computation and time, then step by step digging deeper into the page to extract the data. Therefore the resources of the server or devices, as well the users resources such as time, are saved when there is no need to run the DPI around the clock, nor a need to load the pages entire content for analysis

3.6.2 Maintainability

Any changes in requirements are changed in the system accordingly with minimal changes. As reflected in MVC and other design patterns used.

3.7 Operational Scenarios

3.7.1 Use Case Diagram

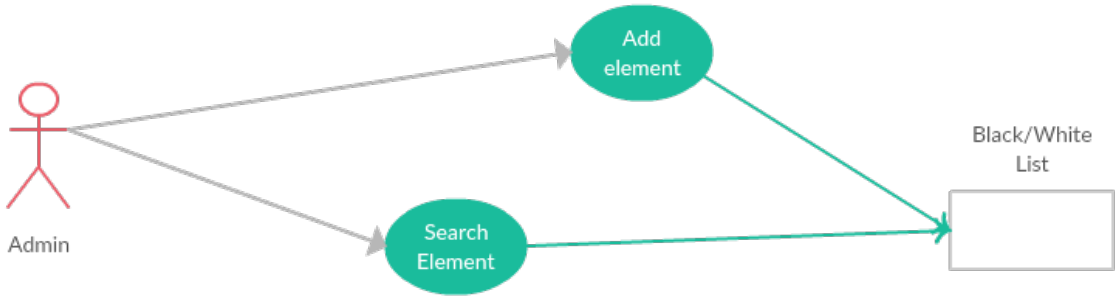


Figure 3.2: Use Case diagram, part 1

Admin of this system can update the black list used by the firewall not only the system could do this

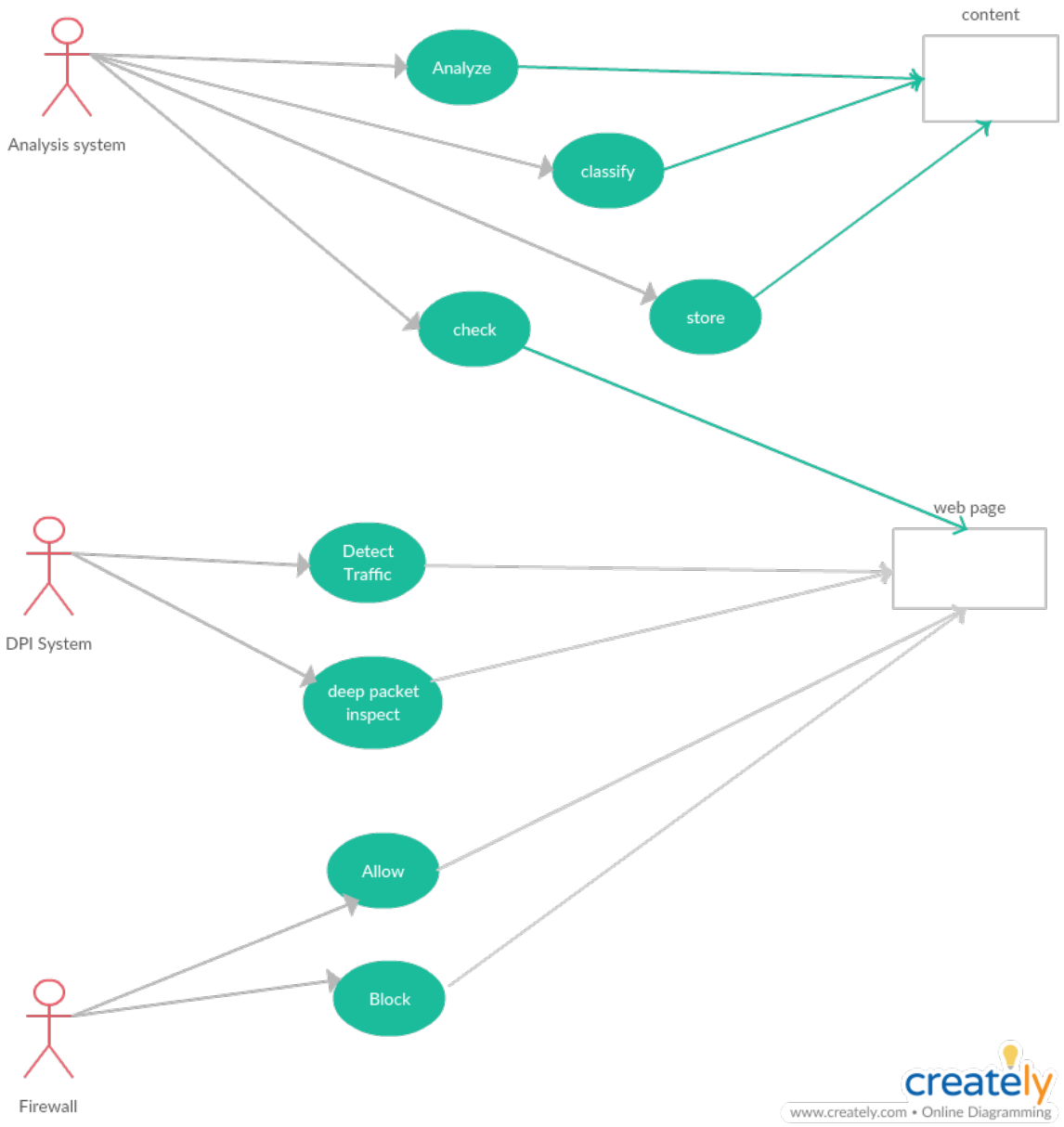


Figure 3.3: Use Case diagram, part 2

the DPI system detects traffic and do deep packet inspection on it, then store it on file that is analyzed and store as a record in a CSV file and the this recored classified and if the class is not education then this URL will be added to the black list and blocked

3.7.1.1 Main Scenario

1. Detect Traffic: running in the background to detect any incoming network traffic.
2. Check: The Analysis system then checks if it already exists. If the Check shows that the web page's information is already classified and stored, then it will take the action.
3. Deep Packet Inspect: To reach the content of the web page to be analyzed, Deep Packet Inspection is needed.
4. Analyze: The received unstructured content is analyzed and turned into structured text to be classified.
5. Classify: classifies the structured content into its equivalent class or category.
6. Store: Storing the web page's meta-data and its corresponding class for future reference.
7. Block: the Firewall blocks the web page according to its class.
8. Allow: the Firewall allows the web page according to its class.

Chapter 4

Software Design Document

4.1 Introduction

4.1.1 Purpose

The purpose of this Software Design Document is to provide a description of the design of the NET-DPI system fully enough to allow for software development to proceed with an understanding of what is to be built and how it is expected to built. The Software Design Document provides information necessary to provide description of the details for the software and system to be built.

4.1.2 Definitions and Acronyms

1. DPI: DPI stands for Deep Packet Inspection which is an advanced method of examining and managing network traffic. It is a form of packet filtering that locates, identifies, classifies, reroutes or blocks packets with specific data or code payloads that conventional packet filtering, which examines only packet headers, cannot detect.
2. Network Interface Card (NIC) is a circuit board or card that is installed in a computer so that it can be connected to a network. A network interface card provides the computer with a dedicated, full-time connection to a network.
3. URL: Uniform Resource Locator (web address)

4. MVC: MVC stands for Model View Controller, which is a software architectural pattern for implementing user interfaces on computers. It divides a given application into three interconnected parts.
5. Design Patterns: Design Pattern is a general reusable solution to a commonly occurring problem in software design. A design pattern is not a finished design that can be transformed directly into code. It is a description or template for how to solve a problem that can be used in many different situations.e.g strategy, MVC, Decorative.
6. SVM: SVM stands for Support Vector Machine, which is a classifier used in classifying the keywords obtained from the Analysis phase.
7. KNN: K Nearest Neighbors.
8. DT: Decision Tree.
9. Tcpdump : Tcpdump stands for a dump traffic on a network. Tcpdump is a library driven from Wireshark. It used for catching and filtering packets, also having the property of showing the payload of the packets.
10. Classification: Classification is considered an instance of supervised learning, learning where a training set of correctly identified observations is available. An algorithm that implements classification, especially in a concrete implementation, is known as a classifier.
11. NLTK: Natural Language Toolkit.
12. RE: Regular Expression.

4.2 System Architecture

4.2.1 Architectural Design

There are many ways to viewing the system's architecture. This section explains the architecture designs of the system from different aspects. According to Sommerville [15], there are architecture design models, and some of them are referenced here.

4.2.1.1 Hardware Architecture

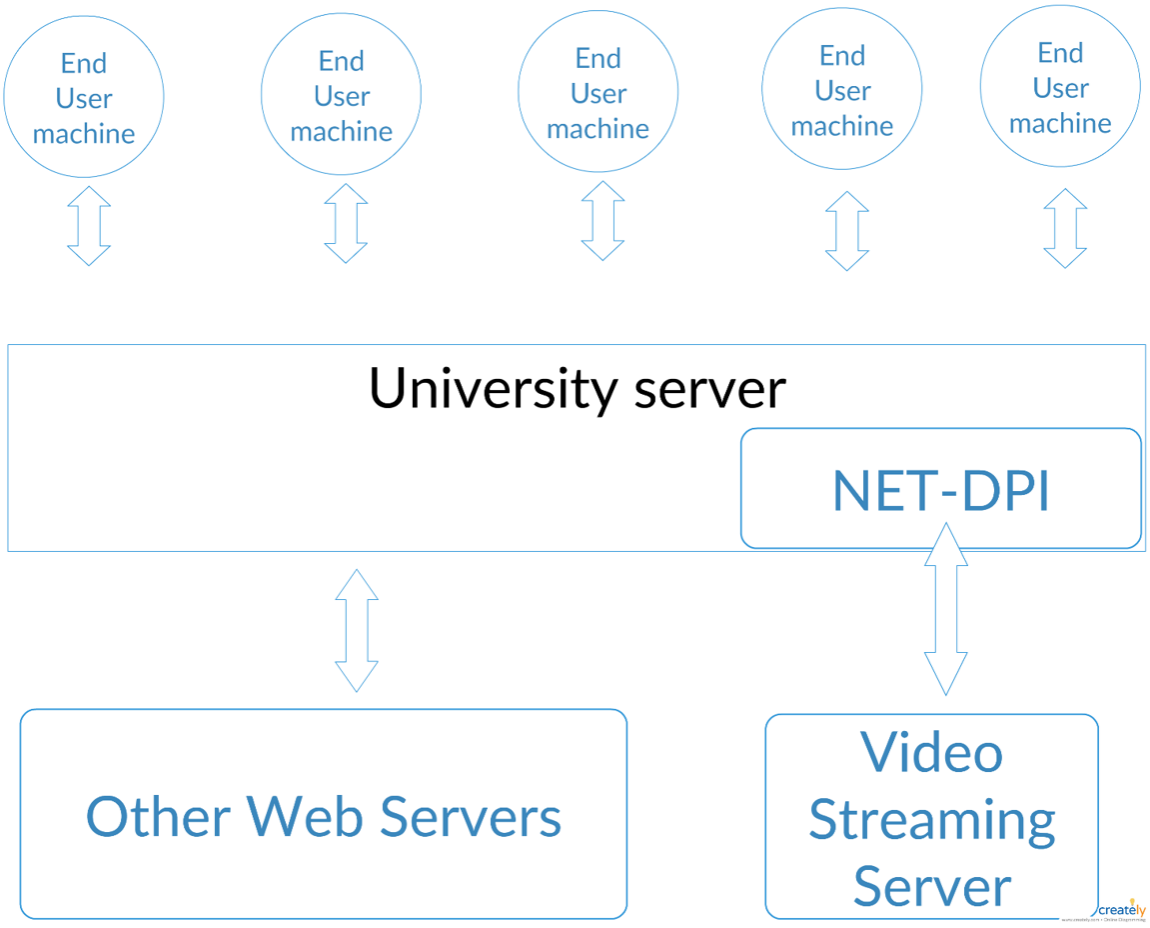


Figure 4.1: Hardware Architecture

The system's hardware architecture in is the view of the system's and external interacting hardware components. The design model used for this architecture is the Client-Server Architecture Design Model, due to the Client-Server nature of most Networks.

4.2.1.2 Process Architecture

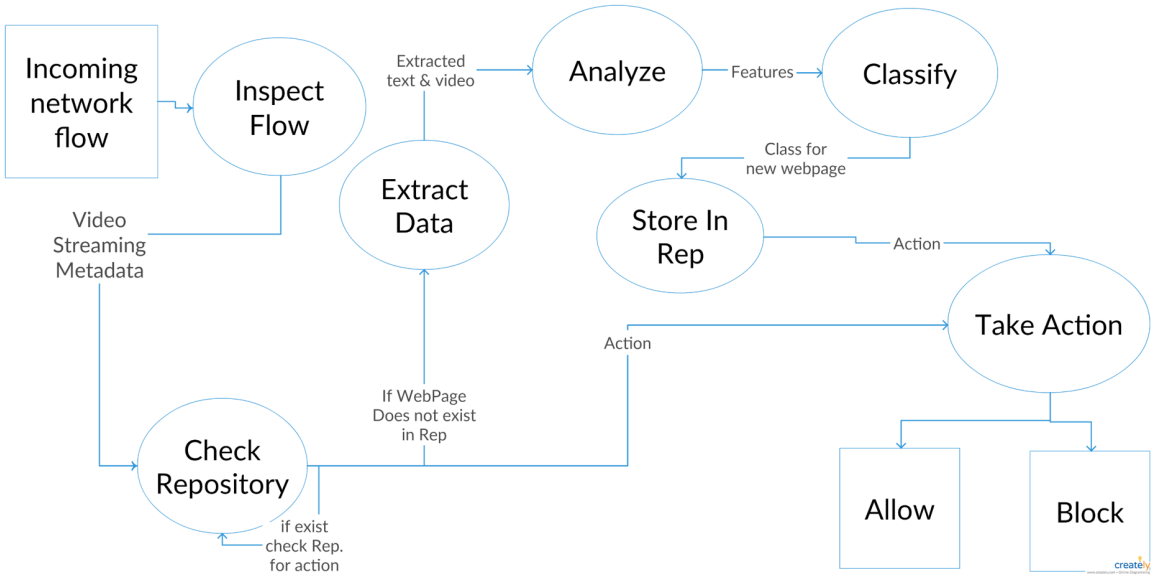


Figure 4.2: Process Architecture

The system's business process architecture is the view of the system's functionality in terms of input, trajectory, and output. The design model used for this architecture is the Pipe-and-Filter Architecture Design Model, to explain the flow of the process. The process shown here is the system's main functionality.

4.2.2 Decomposition Description

4.2.2.1 Class Diagram

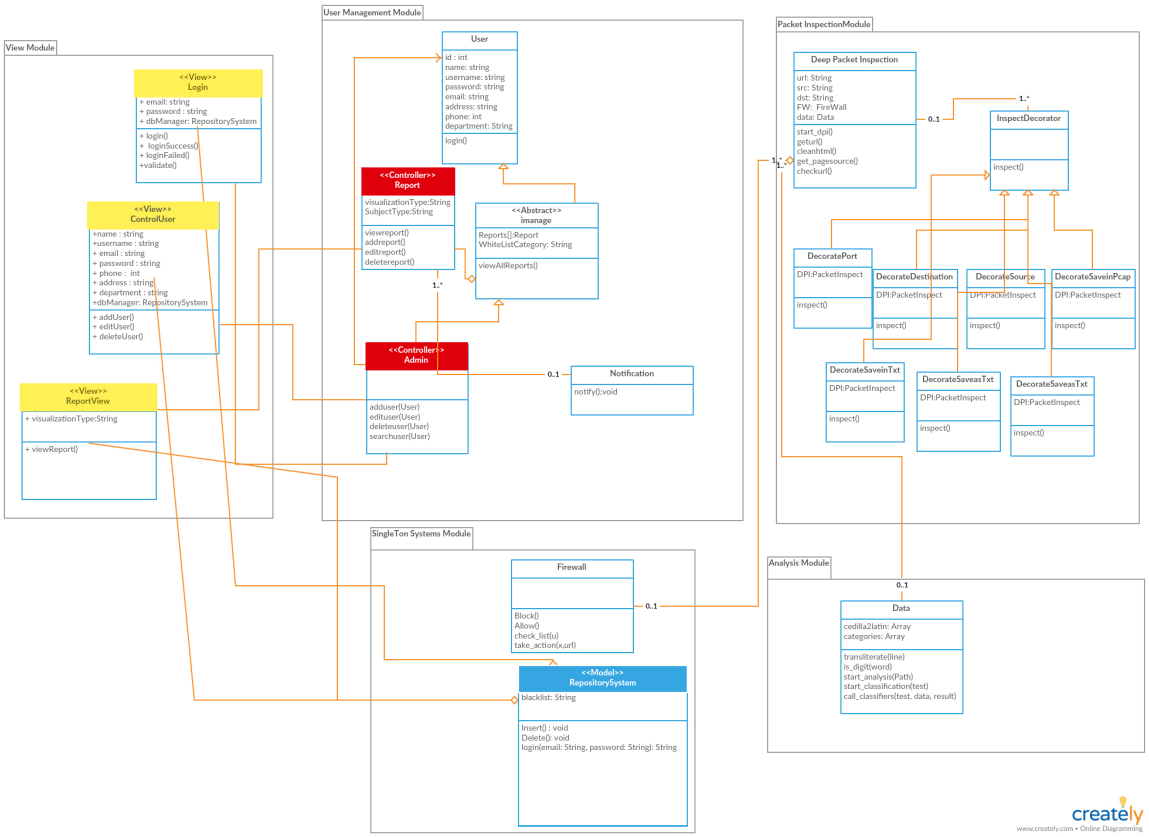


Figure 4.3: Class Diagram

4.2.3 Design Rationale

Hardware Architecture, As the Client-Server is a nature of most Networks. NET-DPI is a network server that placed in the company server to deal with the video streaming servers to make use of the streaming website instead of just blocking it.

Process Architecture, This diagram is used as NET-DPI is not even close to be an interactive system, Input and Output functionality contained, consists of data flow and transformation and processing done in separate stages to generate detailed output.

The library tcpdump [16] is used to create "dumps" or "traces" of network traffic. It allows you to look at what is happening on the network and really can be useful for troubleshooting many types of issues including issues that aren't due to network communications. Outside of network issues I use tcpdump to troubleshoot application issues all the time; if you ever have two applications that don't seem to be working well together, tcpdump is a great way to see what is happening. This is especially true if the traffic is not encrypted as tcpdump can be used to capture and read packet data as well.

K-nearest neighbors (KNN) is widely used classification technique. It is commonly used for its easy of interpretation and low calculation time. The choice of the parameter k is very crucial in this algorithm. The training error rate and the validation error rate are two parameters which are needed to be accessed on different k value.[17]

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Gradient Boosting builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage n classes regression trees are fit on the negative gradient of the binomial or multinomial deviance loss function. Binary classification is a special case where only a single regression tree is induced.

SquidGuard is a combined filter, redirector and access controller plugin for Squid.

4.2.4 Data Design

4.2.4.1 Data Description

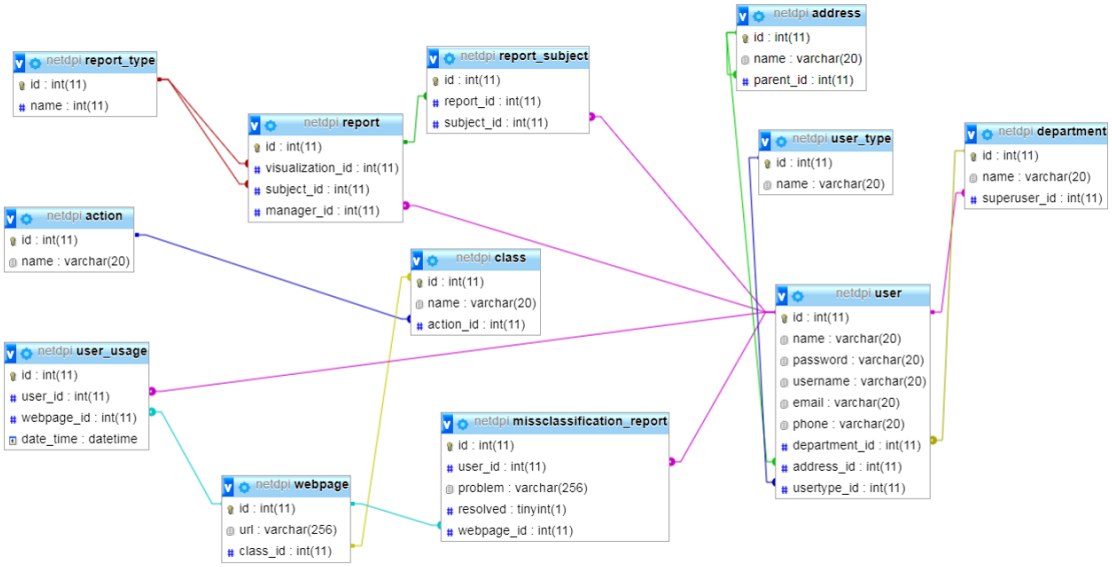


Figure 4.4: Database

4.2.4.2 Data Dictionary

- **User:** This collection has the details of the controllers of the of NET-DPI
- **Blacklist:** This collection has the details of the URLs that have been blocked by the system.
- **Action:** This collection has the details whether the system decided to block or allow the URL.

4.3 Component Design

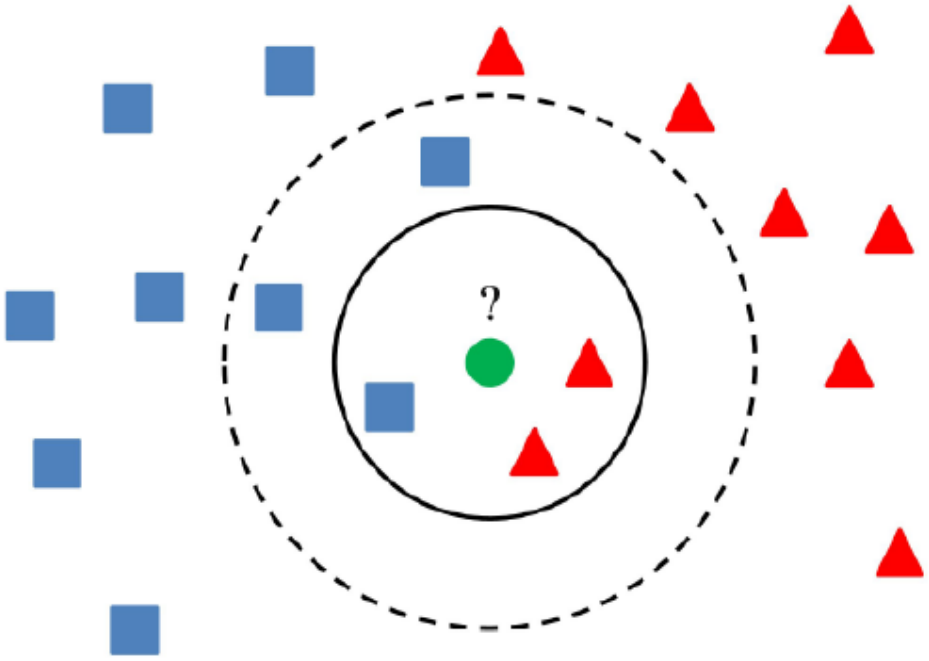


Figure 4.5: KNN

K-nearest neighbors (KNN) is widely used classification technique. It is commonly used for its easy of interpretation and low calculation time. The choice of the parameter k is very crucial in this algorithm. The training error rate and the validation error rate are two parameters which are needed to be accessed on different k value.[17]

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

For instance, in figure 4.6, decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model.

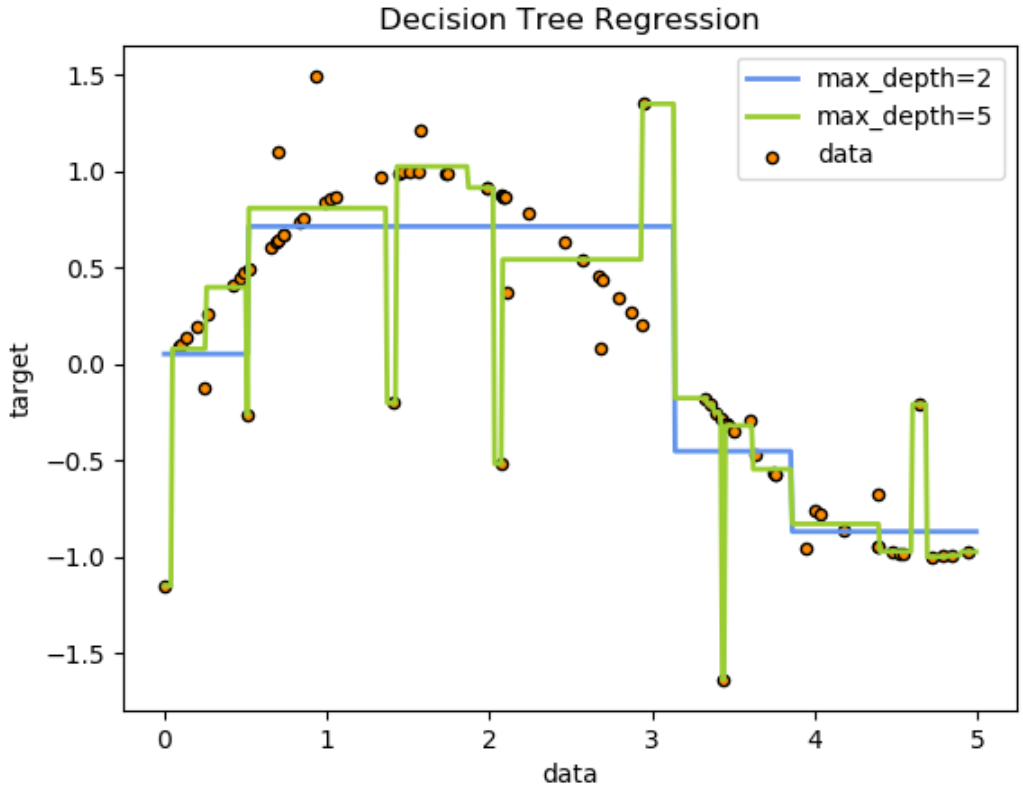


Figure 4.6: Decision Tree

Gradient Boosting builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage n classes regression trees are fit on the negative gradient of the binomial or multinomial deviance loss function. Binary classification is a special case where only a single regression tree is induced.

Tcpdump takes input network traffic and output the stream of packets of certain transactions with text format or hexadecimal format. the pre processing is the decision of what kind of inspection will take place ex(text format, hexadecimal format, inspect from specific IP or port number). the post processing is to take the output and pass it to analyze the content.

SquidGuard is a combined filter, redirector and access controller plugin for Squid.

4.4 Human Interface Design

4.4.1 Overview of User Interface

Our user interface is a web application because our system works on university or company server. So the web application to let the users have some control on the system. The admin is the one who has the total control. The system can generate reports of monitored users. In addition, the admin can add, edit or delete any user. Our system work in the background so it is not visible what it is doing.

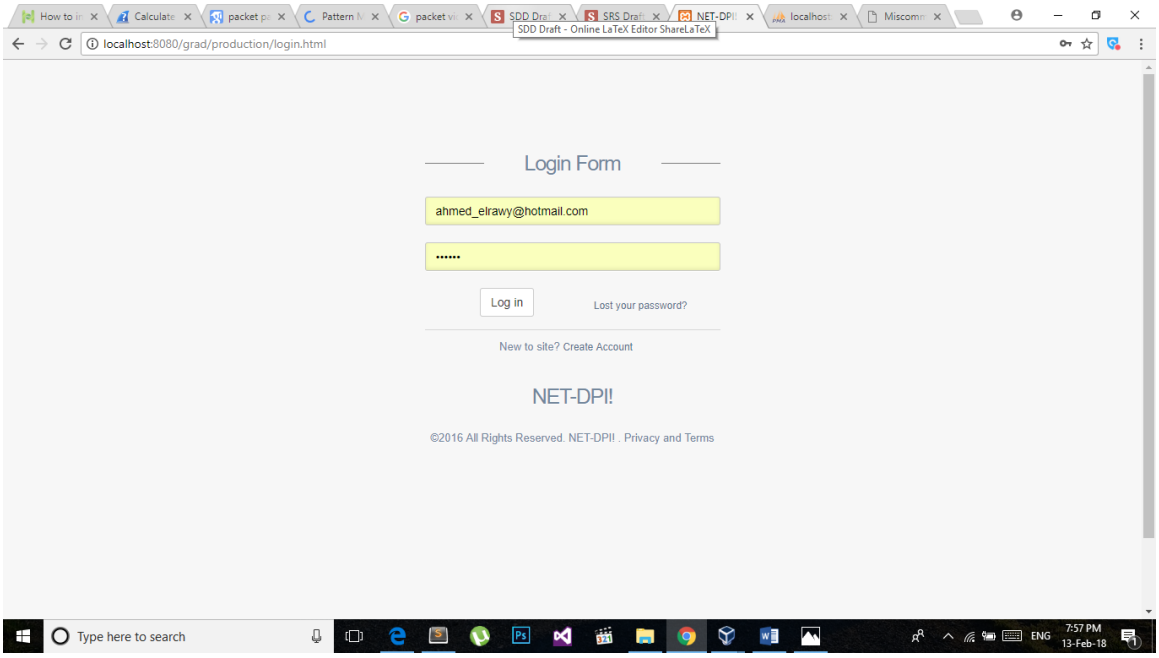


Figure 4.7: Login

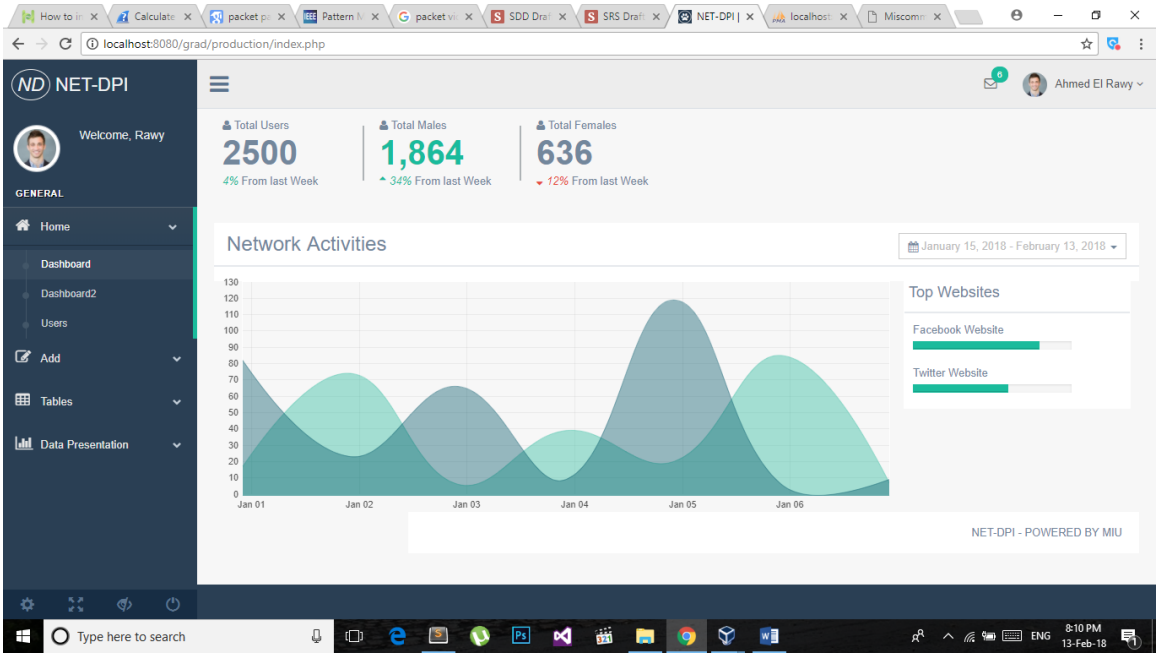


Figure 4.8: Reports 1

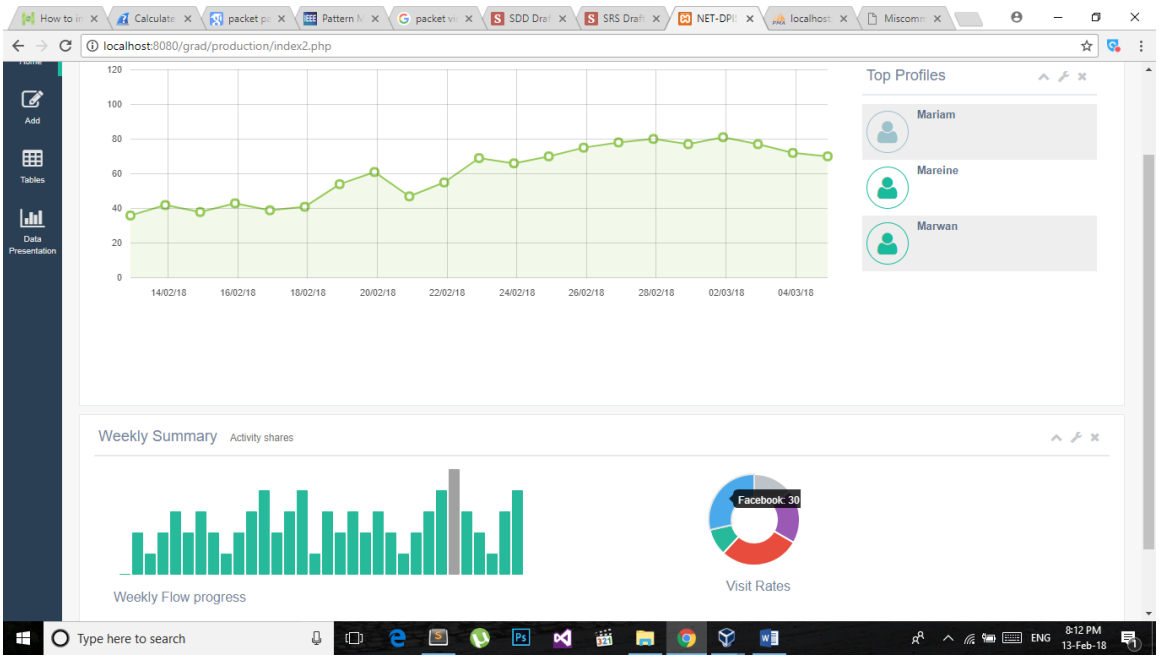


Figure 4.9: Reports 2

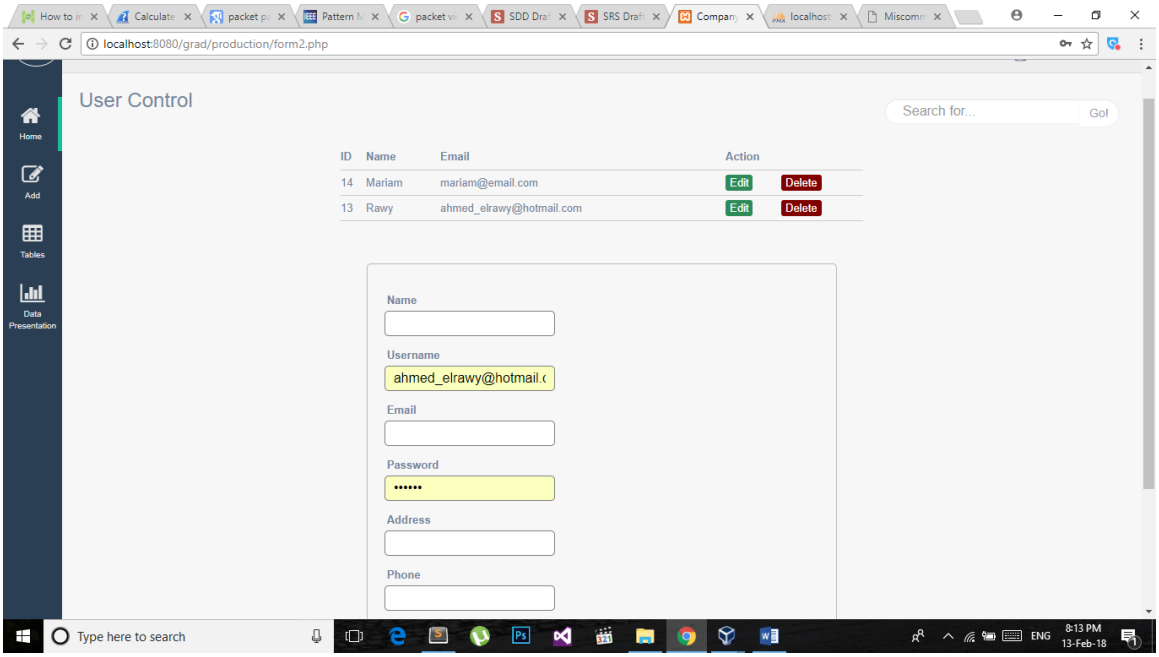


Figure 4.10: User Control

4.4.2 Screen Objects and Actions

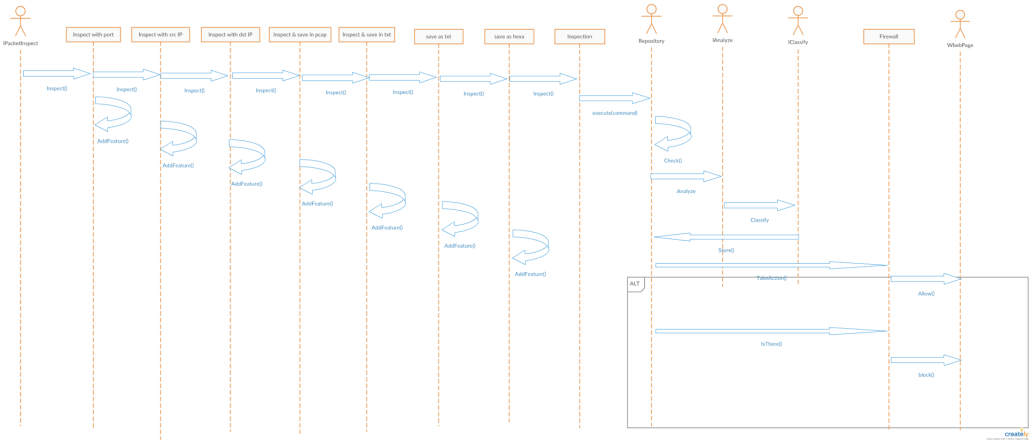


Figure 4.11: Sequence Diagram

4.5 Requirements Matrix

Req. Id	Req. Type	Req. Name	Req. Description	Module	Reference	Status
F01	Functional	Login	Admin Login to System.	User Management	Class Diagram	Completed
F02	Functional	Control User	Add, Edit and delete users	User Management	Class Diagram	Completed
F03	Functional	Inspect	Deep packet inspection for network flow.	Packet Inspection	Class Diagram	Completed
F04	Functional	Execute	Execute Commands to specify what function to fire.	Packet Inspection	Class Diagram	Completed
F05	Functional	TextAnalysis	Analyze the output of the text payload.	Analysis	Class Diagram	Completed
F06	Functional	TextClassification	Classify the output of the textAnalysis.	Analysis	Class Diagram	Completed
F07	Functional	Allow	Allow users to access specific content of website.	Firewall	Class Diagram	Completed
F08	Functional	Block	Block users to access specific content of website.	Firewall	Class Diagram	Completed

Figure 4.12: Requirements Matrix

Chapter 5

Evaluation of the proposed project

5.1 Introduction

After the systems main functionalities and blocks have been implemented, the NET-DPI system passes through the evaluation phase. In this phase, the NET-DPI is evaluated via several experiments that were designed to test the system. The testing has been divided into various experiments. Briefly the experiments are as follows(further details of each experiment is explained in the coming sections of this chapter):

1. The first experiment is done to experiment with different classification algorithms to choose the most fitting to the system.
2. The second experiment is conducted after the classification algorithm is chosen: the fitting of the classification as one of the modules of the system.
3. The third experiment is the User Study of the system.

5.2 Experiment 1: Algorithms

5.2.1 Setup

The construction of the datasets (whether the 4k or the 148k) was considered new and not seen before. Furthermore, as for the type of these datasets it is in text. Now in order to grow the dataset; several stages must be done first. First stage, a text file is created to save in it URLs, each of it's rows were in the shape of URL@category, where these categories are the 15 YouTube categories as well as the URLS are from YouTube. Secondly, each of

the these rows (URL@Category) have their page sources extracted to be used which is the output of this stage. Consequently the output of the previous phase is used as an input for the Analysis phase, where the process of text mining is performed on it, whether having bag of words that are categorized with YouTube categories, as well as using regular expression to clean the page source. Afterwards, Feature Extraction is used to transform the output of the previous stage into numbers, these numbers are counter to it's corresponding category to state how many times was this category mentioned. Finally the output of the previous phase is the final stage that is shown and used in these datasets. During the course of this project, their were two datasets, one 4k and after increasing in it's size and giving each categories equal amount it enlarged to 148k. At the end, the reason their was tremendous differences between the accuracies of the two dataset was the over-fitting since the categories were not evenly distributed in the 4k dataset.

Finally, the algorithms that will be tested on are K-Nearest Neighbor (KNN) which is used for classification and regression with input that consists of the k closest training examples according to Euclidean Distance[18], Decision Tree is a tool that uses a tree-like graph of decisions and their possible consequences.[19], as well as Naive Bayes which is one of the top 10 data mining algorithms due to its simplicity, efficiency, and efficacy. [20] , Support Vector Machine (SVM) that achieves good performance when applied to real problems, especially text-categorization problems. [21], Gradient Boosting achieves state-of-the-art performance in academia, industry, and data analytics competitions. [22], Recurrent Neural Network (RNN) which is a special type of neural network equipped with additional recurrent connections.[23], and finally Convolutional Neural Network (CNN) which is a simple and fast algorithm, it introduces a new way to do unsupervised feature learning, and it provides discriminative features which generalize well. [24].

5.2.2 Goal

The goal of this experiment to find the best classification algorithm or combination of algorithms best suited for NET-DPI's needs.

5.2.3 Results

Table 5.1: Classification Accuracy Results

Classification Algorithm	**data set Size Variance**	
	with 4K	***with 148K***
KNN	100%	99%
Decision Tree	97 %	95%
Naive Bayes	22%	85%
SVM	94%	92%
Gradient Boosting	99%	94%
RNN	90%	92%
CNN	94%	93%

5.2.4 Discussion

For every result there is an explanation. Whether the result was positive or negative. The following are the explanations behind the results of this experiment.

5.2.4.1 Neural Network

The reason why the accuracies of the neural network algorithms whether CNN or RNN we're not as high as the other classification algorithms is that it requires a huge dataset.

5.2.4.2 Naive Bayes

Naive Bayes was tested on both datasets(4k,148k), the reason their is a difference in accuracy between the two datasets was quite enormous is that over fitting acquired. And that due to the fact the URLS for educational in the first dataset (4k) was more than the other URLS in the 4k dataset.

5.2.4.3 Naive Bayes, SVM

These classifiers didn't came out with the best accuracies and that is because these classifiers come out with the best accuracies when they are dealing with text. And since these two datasets are in numbers (because of the feature extraction stage) these classifiers

didn't come out with the best accuracies.

5.2.4.4 Conclusion

In conclusion, these three algorithms came out with the best accuracies due to the fact these algorithms deal with numbers better. And since the two datasets were in numbers because the as mentioned in the setup the text was transformed into numbers after applying feature extraction. Furthermore, their complexities are less than the other for faster computation.

5.3 Experiment 2: Applying of System Modules

5.3.1 Setup

NET-DPI is made of three main modules: Firewall, DPI, and Data Analysis and Classification. The Classification algorithms to be used have been determined through the previous experiment. The Firewall module is based on the SquidGuard firewall; the firewall is prepared to block or allow based on the Classification output. The DPI module is based on the tcpdump library, which is for the capturing of the packets in a network traffic. As a setup, the three modules are working separately and statically.

5.3.2 Result

Many bugs were found and fixed. However one major challenge was presented: The speed of the processing needs to be faster than the web-page's loading time. The system is now ready for user's test and use. Finally, finding the domains and protocols that can be captured, and adjusting the system to be equipped with handling different types.

5.3.3 Discussion

The modules of the system needed to be integrated in a certain order. It would be found out that the firewall will run from the server at all times even without the system's cycle launch, so the previous classifications of URL's would be applied in the form of block or allow accordingly, continuously. The capturing of the network's traffic would be first in

order for the cycle to start, and accordingly a series of cascading decisions are taken by the system to determine the course of action.

5.3.3.1 DPI Module

The DPI would receive the capturing of the tcpdump, and firstly extract the URL of the network transaction from the capture. The URL could be in one of two things: within the list of domains supported by NET-DPI, or out of its scope. If the URL is within scope, then another check is made to see if it is the type of URL that accompanies the page-source in the tcpdump capture, or if the tcpdump could only capture the URL. If the capture contains the page-source, then some cleaning of the capture is required to rid it of noise (other packets around the transaction). However if only the URL was captured, the page-source is requested by that URL.

5.3.3.2 Firewall Module

After the DPI is finished with its URL analysis, the URL is sent to the Firewall module to check its list if the URL already exists. If the URL exists in the list, then the firewall blocks or allows accordingly.
After classification, the allow or block decision is made according to the classification's output.

5.3.3.3 Data Module

The output is turned into True or False so it could be directly linked to the firewall.

5.4 Experiment 3: User Study

5.4.1 Setup

In order to pave the way for NET-DPI to be tested on regular users; several things had to be done first. First of all, 4 websites were found that were HTTP, these 4 websites were: First website was an university website, second websites were two websites that contained courses and their materials as well as research materials, the third websites were two websites HTTP non-secured video streaming websites, and the fourth was a entertainment website. The reason NET-DPI currently can only be tested on HTTP is that testing on HTTPs will not provide it's page source because it is encrypted. Furthermore, to simulate this experiment; the user run the system on a computer, while the program was on server. Finally, after applying the previous steps; regular users were able to participate in this experiment, trying out the system.

5.4.2 Goal

The goal of this experiment is to simulate that NET-DPI is applicable and working in real life. As well as the users are satisfied with the end product.

5.4.3 Discussion

Users came into contact with system after testing it on the websites mentioned above. Each user came out with different experience as well as different results. The system was tested on four users. The first user tested 4 URLs, the first URL was educational website and was classified as educational web page and that occurred because their was a matching between the website's text content and NET-DPI bag of words.The second website was an entertainment website which was classified as noneducational website and that happened because the content of this website contained words that were considered under the entertainment category. And the third website was an educational website and it was misclassified and that happened because the three classifiers that are used all of them must outputted that same result which is either true or false classification, and in this case two of the three classifiers predicted that their result is true (education) while the third classifier predicted it is false(non-education),so the final result came out to be non-education. And the last website was a HTTP video streaming website, the video that was captured was a

sports video and output of the classifiers were non-educational web page.
The second user's experience was different than the first user. This user tested 4 URLs each URL had different category. The first URL was an university website, and the output of the three classifiers was that this web page was a non-educational web page as it contains portfolio and news about the university new comers and other related posts and it does not contain any information about any courses and their materials. The second website was an entertainment website and it was classified by the three classifiers as a non educational web page which is considered classified right. The third URL was an HTTP video streaming website, the video that was captured it's title resembled an that this video was an educational video while it's descriptions and comments were not an educational. It was classified as a non- educational content. The fourth website was as the previous URL which is an HTTP video streaming website, the video was about car racing, and it was classified as non-educational web page.

The third user tested 4 URLs, the first URL was an educational website and the output of the three classifiers was that this web page educational web page and that occurred because their was a matching between the website's text content and NET-DPI bag of words. The second URL was an entertainment website, and it was classified as a non-educational web page and that because the content of this page was entertainment related. The third URL was an HTTP video streaming website, the video that was tested on was educationally related, and it was successfully classified as an educational web page. The fourth URL was as the previous an HTTP video streaming website, as for the video's content was educationally related, and it was classified correctly as education web page.

The forth user's experience's result was like the second user but with different URLs. This user tested 4 URLs each URL had different category. The first URL was an university website, and the output of the three classifiers was that this web page was a non-educational web page and that happened because as mentioned above this web page does not contain any material nor information about courses, only news about the events the university is participated in. The second website was an entertainment website, and the result of the three classifiers came out that this web page was an non-educational web page, and that happen because this web page contained words that are demonstrated as entertainment. The third URL was an HTTP video streaming website, the video's content was beauty

related, and the classifiers outputted that this web page is non-educational, and that due to the fact that the tittle of the video as well as the description and the comments are beauty related. The fourth of URL was an educational website, but it was misclassified, and that happened because this web page had only information about the team members of this education website and had no educational information.

5.4.4 Results

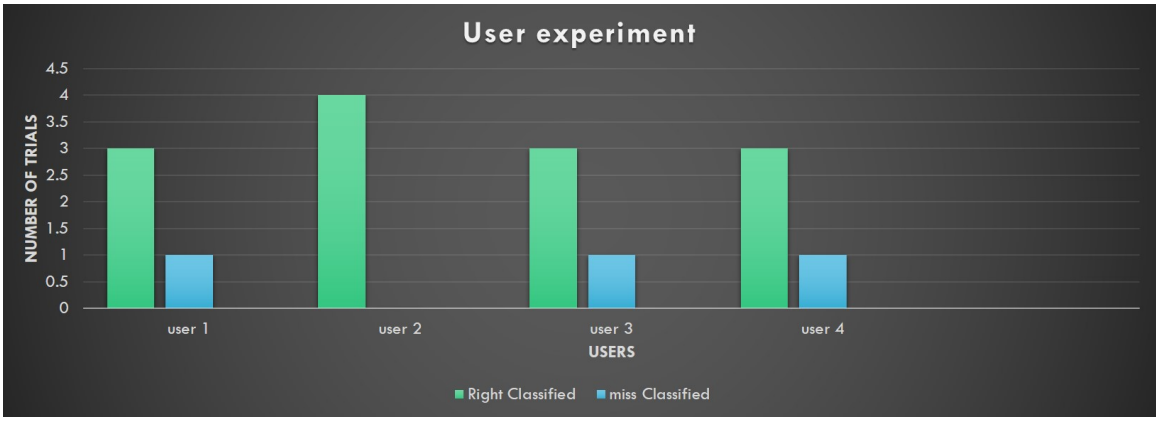


Figure 5.1: Experiment 3 Results

Chapter 6

Conclusion

Throughout the process of implementing NET-DPI, some challenges were faced. One of these challenges was enhancing the dataset to in order to improve the accuracies of Neural Network classifiers. As well as most of the transactions can only be read at the destination in this case the client, if the computer acts as a server without the client's permission then the server can't access that data, which causes the traffic to be encrypted. Furthermore, running NET-DPI at it's full power and speed requires server and super computer which is currently not available. In addition to not having several user types defeats the purpose of generating reports about them.

6.1 Future directions

In the future, the majority of the challenges mentioned above can be solved by integrating NET-DPI with a standard firewall or network filter system.This standard system has already users with their corresponding IDs as well as user types (such as: employers, employees, head of departments..etc). With that this standard system solves the challenges face above such as encrypted traffic problem because the user log in on the network giving NET-DPI the permission to access his/her data. As well as the challenge of user identification to help NET-DPI with generating reports about the user's usage of the network. Furthermore, this standard system is equipped with handling multiple clients at the same time. Finally running NET-DPI on server computer, speed will not be an issue.

Bibliography

[1] B. Anderson and D. Mcgrew, "Machine learning for encrypted malware traffic classification," *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 17*, 2017.

[2] K. Ihalagedara, R. Kithuldeniya, S. Weerasekara, and S. Deegalla, "Feasibility of using machine learning to access control in squid proxy server," *2015 IEEE 10th International Conference on Industrial and Information Systems (ICIIS)*, 2015.

[3] "telecomhall." [Online]. Available: http://www.telecomhall.com/

[4] "Tm forum inform." [Online]. Available: https://inform.tmforum.org/

[5] S. Ansari, S. G. Rajeev, and H. S. Chandrashekar, "Packet sniffing: a brief introduction," *IEEE Potentials*, vol. 21, no. 5, pp. 17–19, Dec 2002.

[6] H. Zimmermann, "Osi reference model - the iso model of architecture for open systems interconnection," *IEEE Transactions on Communications*, vol. 28, no. 4, pp. 425–432, April 1980.

[7] T. J. Parvat and P. Chandra, "Performance improvement of deep packet inspection for intrusion detection," in *2014 IEEE Global Conference on Wireless Computing Networking (GCWCN)*, Dec 2014, pp. 224–228.

[8] N. Boudriga, *Security of mobile communications.* CRC Press/Taylor Francis, 2010.

[9] T. R. Peltier and J. Peltier, *Complete guide to CISM certification.* Auerbach, 2007.

[10] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, p. 147, Jan 2002.

[11] M. Mishra and M. Srivastava, "A view of artificial neural network," in *2014 International Conference on Advances in Engineering Technology Research (ICAETR - 2014)*, Aug 2014, pp. 1–3.

[12] M. Lotfollahi, R. S. H. Zade, M. J. Siavoshani, and M. Saberian, "Deep packet: A novel approach for encrypted traffic classification using deep learning," *CoRR*, vol. 1709.02656, 2017.

[13] H. Yu, Y. Zhao, G. Xiong, L. Guo, Z. Li, and Y. Wang, "Poster: Mining elephant applications in unknown traffic by service clustering," *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS 14*, 2014.

[14] A. Bremler-Barr, S. T. David, Y. Harchol, and D. Hay, "Leveraging traffic repetitions for high-speed deep packet inspection," *2015 IEEE Conference on Computer Communications (INFOCOM)*, 2015.

[15] I. Sommerville, *Software engineering.* Addison-Wesley, 2016.

[16] Tcpdump, "Tcpdump/libpcap public repository," 2010. [Online]. Available: https://www.tcpdump.org/

[17] A. Moldagulova and R. Sulaiman, "Using knn algorithm for classification of textual documents - ieee conference publication," Oct 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8079924/

[18] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, p. 175, 1992.

[19] J. Quinlan, "Simplifying decision trees," *International Journal of Human-Computer Studies*, vol. 51, no. 2, p. 497510, 1999.

[20] C. Yan, "Innovative applications of artificial intelligence," *Engineering Applications of Artificial Intelligence*, vol. 3, no. 2, p. 166167, 1990.

[21] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, July 1998.

[22] J. Jiang, J. Jiang, B. Cui, and C. Zhang, "Tencentboost: A gradient boosting tree system with parameter server," in *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, April 2017, pp. 281–284.

[23] B. Li, E. Zhou, B. Huang, J. Duan, Y. Wang, N. Xu, J. Zhang, and H. Yang, "Large scale recurrent neural network on gpu," in *2014 International Joint Conference on Neural Networks (IJCNN)*, July 2014, pp. 4062–4069.

[24] A. Ghaderi and V. Athitsos, "Selective unsupervised feature learning with convolutional neural network (s-cnn)," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Dec 2016, pp. 2486–2490.